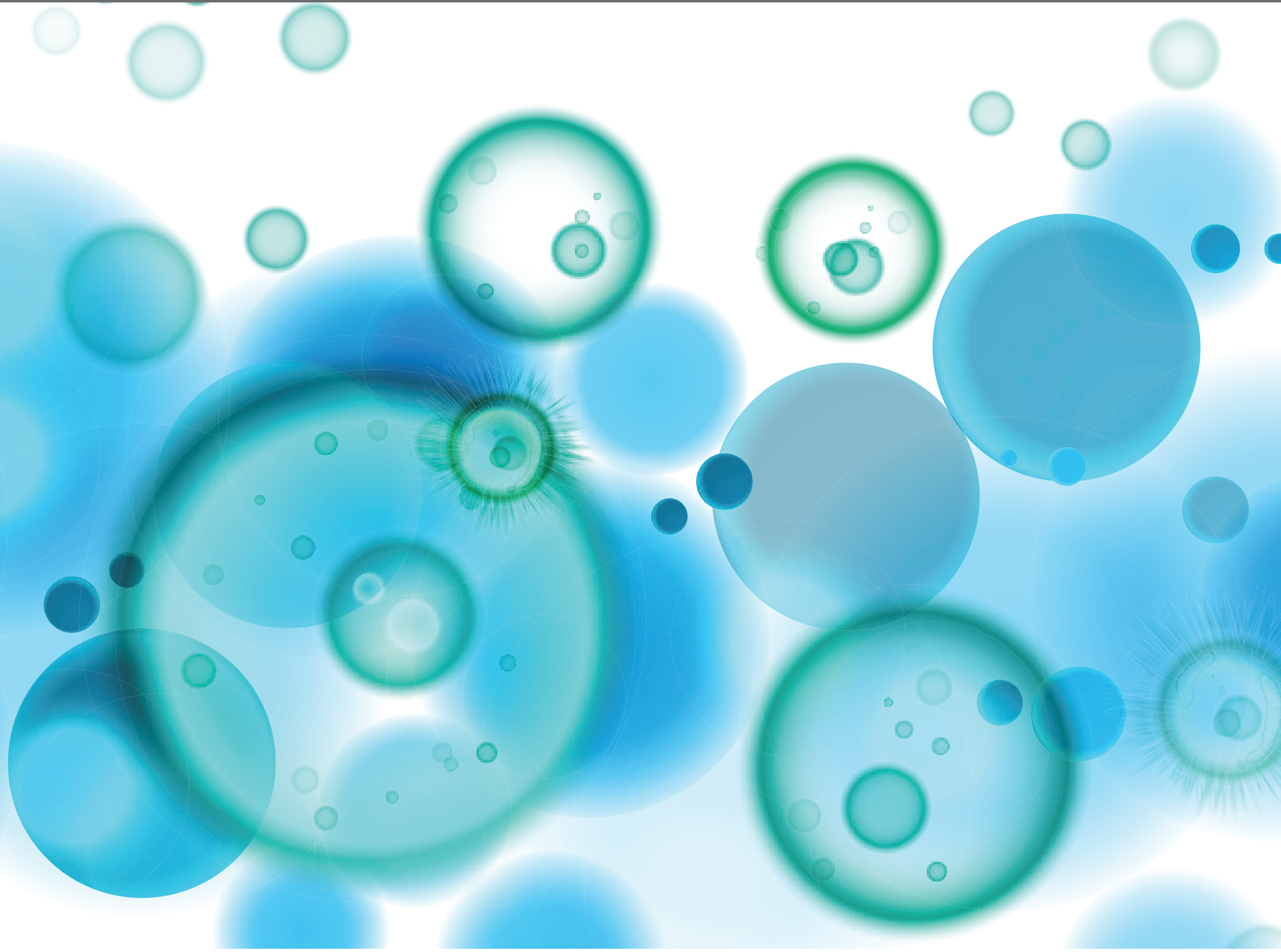
# HLA AND KIR DIVERSITY AND POLYMORPHISMS: EMERGING CONCEPTS

EDITED BY: Malini Raghavan, Ramit Mehr, Yoram Louzoun, Martin Maiers and Jim Kaufman

PUBLISHED IN: Frontiers in Immunology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-036-2
DOI 10.3389/978-2-88971-036-2

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# HLA AND KIR DIVERSITY AND POLYMORPHISMS: EMERGING CONCEPTS

Topic Editors:
**Malini Raghavan,** University of Michigan, United States
**Ramit Mehr,** Bar-Ilan University, Israel
**Yoram Louzoun,** Bar-Ilan University, Israel
**Martin Maiers,** National Marrow Donor Program, United States
**Jim Kaufman,** University of Cambridge, United Kingdom

**Citation:** Raghavan, M., Mehr, R., Louzoun, Y., Maiers, M., Kaufman, J., eds. (2021).
HLA and KIR Diversity and Polymorphisms: Emerging Concepts.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-036-2

# Table of Contents

frontiers in Immunology

EDITORIAL
published: 17 May 2021
doi: 10.3389/fimmu.2021.701398

# Editorial: HLA and KIR Diversity and Polymorphisms: Emerging Concepts

*Martin Maiers*[1*], *Ramit Mehr*[2], *Malini Raghavan*[3], *Jim Kaufman*[4,5] *and Yoram Louzoun*[2*]

[1] National Marrow Donor Program, Minneapolis, MN, United States, [2] Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, [3] Michigan Medicine, University of Michigan, Ann Arbor, MI, United States, [4] Department of Pathology, University of Cambridge, Cambridge, United Kingdom, [5] Institute for Immunology and Infection Research, University of Edinburgh, Edinburgh, United Kingdom

**Keywords: HLA, KIR, diversity, immunogenetics, polymorphism SIP- Society for Immune Polymorphism**

**Editorial on the Research Topic**

**HLA and KIR Diversity and Polymorphisms: Emerging Concepts**

OPEN ACCESS

***Edited and reviewed by:***
*Antoine Toubert,*
*Université Paris Diderot, France*

****Correspondence:***
*Martin Maiers*
*mmaiers@nmdp.org*
*Yoram Louzoun*
*louzouy@math.biu.ac.il*

***Specialty section:***
*This article was submitted to Alloimmunity and Transplantation, a section of the journal Frontiers in Immunology*

***Received:*** *27 April 2021*
***Accepted:*** *30 April 2021*
***Published:*** *17 May 2021*

***Citation:***
*Maiers M, Mehr R, Raghavan M, Kaufman J and Louzoun Y (2021) Editorial: HLA and KIR Diversity and Polymorphisms: Emerging Concepts. Front. Immunol. 12.701398. doi: 10.3389/fimmu.2021.701398*

Polymorphisms of immune system genes, including Human Leukocyte Antigens (HLA) and the Killer Ig-Like Receptors (KIR), offer limitless depths of complexity. Jean Dausset, who shared the Nobel Prize in 1980 for the discovery of the HLA system, when presented with new results showing the role of specific amino acid polymorphisms on the function of HLA, responded enthusiastically that "we need to go further and study HLA at the atomic level".

The major histocompatibility complex (MHC) and KIR genetic loci are strongly associated with the outcomes of infectious diseases, cancers, inflammatory and autoimmune diseases, reproduction, and transplantation. Recent advances in characterizing HLA and KIR diversity in human populations have led to translational impacts for bone marrow and solid organ transplant matching.

There is currently robust discussion on the origins of HLA, MHC and KIR polymorphisms in humans, primates and other mammals, but limited information about the effects of the polymorphisms on interactions mediated by HLA and KIR. While both the HLA and KIR immune genes regions have been independently associated with diseases and their outcomes, the interactions between those regions has had limited attention. This Research Topic was conceived to appeal to researchers that have found their way, *via* different paths, to the crossroads of the polymorphisms and population dynamics of a *particular* immune sub-system with the recognition that, to progress in our understanding, we also need to establish a form of systems immunology to study the *interaction* of various polymorphic components.

There are established studies that focus on each of these areas: HLA, T cells, B cells, NK cells and their receptors. We are interested in the intersection: what emerges in combination at the synapses of specific molecules in the context of the entire organism or population.

The resulting Research Topic is 15 papers, 14 of which primarily focus on only one of the systems in humans and primates, and provide important insights into receptor repertoire diversity, diversification mechanisms, haplotypes, expression, sequencing, donor matching, peptide repertoires, and disease linkages among:

-**KIR and other NK receptors** (Solloch et al.; Alicata et al.; Roe, Vierra-Green et al.; Roe, Williams et al.; Roe and Kuang; Bruijnesteijn et al.; Cisneros et al.; Cubero et al.)

-**HLA/MHC** (Yamamoto et al.; Kavadichanda et al.; Nunes et al.; Kaufman**)**

-**MICA/B** (Klussmeier et al.)

-**Immunoglobulin heavy chains** (Rodriguez et al.)

Alas only one study explicitly focused on the interaction of HLA and KIR (Vargas et al.).

The studies presented in these articles have implications for understanding disease risk, outcomes and pathogenic mechanisms, host defense outcomes, gene classification and nomenclature, and transplantation. The studies also identify a gap - the need for new ways to bring together researchers in adjacent areas studying genetic polymorphism in the different branches of the immune system and to expose them to the various challenges, methods and advances in each sub-field.

To that end, we have developed a charter for a new *Society for Immune Polymorphism - SIP*. We envision this society as an international membership organization of scientists dedicated to understanding genetic and functional variations in the vertebrate immune system, and the vertebrate immune system's role in health, disease, and evolutionary biology.

Our goal is to promote increased collaboration, engagement and sharing of domain knowledge between scientists focused on any aspects of immune-related genomics through the following activities

- to provide a forum for basic science, and clinical, industrial, and educational applications related to immune polymorphism.
- to share and disseminate research findings related to immune polymorphism.
- to assist in the integration of immune polymorphism-related data resources.
- to promote the development of secure open-source, cloud-based technologies that advance the analysis, collection, exchange, and storage of immune-related genomic data.

The goals of SIP are not to replace any of the societies interesting in a specific aspect of immune polymorphisms, but rather to allow for fruitful collaborations between practitioners and researchers in the different domains. We have organized two workshops in Ramat Gan, Israel (Mar, 2018 https://louzouy.wixsite.com/hlakir2018/) and in New Orleans (Oct, 2019 https://immunepolymorphism.tulane.edu/). A plan for this type of article collection was conceived at the first workshop. We plan to continue to work together in these efforts and welcome broad participation from a diverse group of researchers around the globe.

The society is still in its initial stages, further details about the society, its goals, and possible collaborations between SIP and other societies can be obtained from the corresponding authors or the website: immunepolymorphismsociety.org.

## AUTHOR CONTRIBUTIONS

MM drafted the original. YL, JK, MR, and RM provided comments and feedback. All authors contributed to the article and approved the submitted version.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

ORIGINAL RESEARCH
published: 18 February 2020
doi: 10.3389/fimmu.2020.00156

# Subordinate Effect of -21M HLA-B Dimorphism on NK Cell Repertoire Diversity and Function in HIV-1 Infected Individuals of African Origin

*Elia Moreno Cubero*[1], *Ane Ogbe*[1], *Isabela Pedroza-Pacheco*[1], *Myron S. Cohen*[2], *Barton F. Haynes*[3], *Persephone Borrow*[1] and *Dimitra Peppa*[1,4]*

*[1] Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom, [2] University of North Carolina School of Medicine, Chapel Hill, NC, United States, [3] Duke University Human Vaccine Institute, Duke University School of Medicine, Durham, NC, United States, [4] Department of HIV, Mortimer Market Centre, Central and North West London NHS Foundation Trust (CNWL), London, United Kingdom*

Natural Killer (NK) cells play an important role in antiviral defense and their potent effector function identifies them as key candidates for immunotherapeutic interventions in chronic viral infections. Their remarkable functional agility is achieved by virtue of a wide array of germline-encoded inhibitory and activating receptors ensuring a self-tolerant and tunable repertoire. NK cell diversity is generated by a combination of factors including genetic determinants and infections/environmental factors, which together shape the NK cell pool and functional potential. Recently a genetic polymorphism at position -21 of HLA-B, which influences the supply of HLA-E binding peptides and availability of HLA-E for recognition by the inhibitory NK cell receptor NKG2A, was shown to have a marked influence on NK cell functionality in healthy human cytomegalovirus (HCMV) seronegative Caucasian individuals. In this study, -21 methionine (M)-expressing alleles supplying HLA-E binding peptides were largely poor ligands for inhibitory killer immunoglobulin-like receptors (KIRs), and a bias to NKG2A-mediated education of functionally-potent NK cells was observed. Here, we investigated the effect of this polymorphism on the phenotype and functional capacity of peripheral blood NK cells in a cohort of 36 African individuals with human immunodeficiency virus type 1 (HIV-1)/HCMV co-infection. A similarly profound influence of dimorphism at position -21 of HLA-B on NK cells was not evident in these subjects. They predominantly expressed African specific HLA-B and -C alleles that contribute a distinct supply of NKG2A and KIR ligands, and these genetic differences were compounded by the marked effect of HIV-1/HCMV co-infection on NK cell differentiation. Together, these factors resulted in a lack of correlation of the HLA-B -21 polymorphism with surface abundance of HLA-E and loss of the NK cell functional advantage in subjects with -21M HLA-B alleles. Instead, our data suggest that during HIV/HCMV co-infection exposure of NK cells to an environment that displays altered HLA-E ligands drives adaptive NKG2C+ NK cell expansions influencing effector responses. Increased efforts to understand how NK cells are functionally calibrated to self-HLA during chronic viral infections will pave the way to developing targeted therapeutic interventions to overcome the current barriers to enhancing immune-based antiviral control.

**Keywords: Natural Killer cells, HIV-1, HCMV, HLA-B, HLA-C**

***Edited by:***
*Ramit Mehr,*
*Bar-Ilan University, Israel*

***Reviewed by:***
*Martin Cornillet,*
*Karolinska Institutet, Sweden*
*Paul J. Norman,*
*University of Colorado Denver,*
*United States*

***Correspondence:***
*Dimitra Peppa*
*dimitra.peppa@ndm.ox.ac.uk*

***Specialty section:***
*This article was submitted to NK and Innate Lymphoid Cell Biology, a section of the journal Frontiers in Immunology*

***Received:*** *16 October 2019*
***Accepted:*** *21 January 2020*
***Published:*** *18 February 2020*

***Citation:***
*Cubero EM, Ogbe A, Pedroza-Pacheco I, Cohen MS, Haynes BF, Borrow P and Peppa D (2020) Subordinate Effect of -21M HLA-B Dimorphism on NK Cell Repertoire Diversity and Function in HIV-1 Infected Individuals of African Origin. Front. Immunol. 11:156. doi: 10.3389/fimmu.2020.00156*

# INTRODUCTION

There is a pressing need to better characterize and harness the immune response in order to develop efficacious immune-based strategies to supplement current therapeutic approaches for a "functional" cure in chronic viral infections. Natural Killer (NK) cells have the potential to respond to viruses as direct effectors and can edit adaptive immunity influencing the outcome of viral infections (1). More recently, their capacity to develop adaptive or memory-like features in the setting of infection has been highlighted (2). A number of studies, both epidemiological and functional, have provided evidence for the important role of NK cells in human immunodeficiency virus type 1 (HIV-1) viral control and protection from acquiring new infection (3).

In order for NK cells to gain functional competence they are required to be "licensed" or educated, a process that refines their levels of responsiveness (4). Traditionally this was ascribed to the presence of inhibitory killer immunoglobulin-like receptor (KIR)—human leukocyte antigen (HLA) class I pairs. However, recent evidence suggests that NK cells can be educated through the older and more conserved inhibitory receptor CD94/NKG2A, which recognizes HLA-E complexed with a peptide derived from the leader sequence of HLA-A, B or C alleles as well as HLA-G (5). HLA-E has little polymorphism and its levels of expression are influenced by peptide ligand availability. Whereas, HLA-A and HLA-C allotypes are fixed for Methionine (-21M), HLA-B contains a polymorphism that can encode either Methionine (-21M), which gives rise to functional HLA-E binding peptides, or Threonine (-21T) at this position, which does not bind effectively to HLA-E. The resultant HLA-B -21M/T variation defines different sets of haplotypes with -21M biasing toward NKG2A NK cell education, which has been shown to be associated with superior NK function in healthy HCMV seronegative adults, and -21T promoting KIR mediated education (6). The reported linkage disequilibrium (LD) in Eurasian populations between HLA-B -21M and HLA-B Bw6/HLA-C1, which interact poorly with KIRs, further decreases their potential to mediate NK cell education through KIR engagement. In contrast, -21T HLA-B haplotypes in various combinations with Bw4, C1, and C2 enhance education via KIRs. Interestingly haplotypes combining HLA-C2 and -21M HLA-B are more frequently found in Africa in combination with HLA-C allotypes that promote HLA-E expression poorly (7). The dimorphism at position -21 of HLA-B (M/M genotype) has been associated with increased susceptibility to HIV-1 infection (8). Notably, the dimorphism influences NK cell cytolysis of HIV-infected CD4 T cells and macrophages *in vitro*, with -21T enhancing cytolysis compared to -21M, suggesting that the more educated NKG2A+ NK cells of M/M donors may be less effective in responding to HIV-1 (9). In light of this, the beneficial effects of Bw4+ HLA-B homozygosity in controlling HIV-1 viraemia could be re-interpreted in terms of a mechanism involving recognition of HLA-E by NKG2A+ NK cells of T/T donors (10). Recently HLA-B haplotypes that favor education via NKG2A were also found to exacerbate the detrimental effect of high HLA-A on HIV-1 control through impaired killing of HIV infected target cells (11). However, this effect of HLA-A expression on HIV-1 viraemia was less pronounced in individuals of African descent, possibly reflecting the distinct frequencies of HLA haplotypes present in these populations (11).

To date, the phenotypic and functional effects of HLA-B -21 dimorphism on NK cells have not been assessed in the context of HCMV seropositive individuals or in HIV-1 infected cohorts, where HCMV co-infection is almost universal (12). We have recently demonstrated the potent effect of HCMV co-infection in shaping the NK cell repertoire during chronic HIV-1 infection, leading to an accelerated differentiation and adaptive reconfiguration of the NK cell compartment and expansion of an NK cell subset expressing NKG2C, the activating counterpart of NKG2A that also binds to HLA-E (recognizing HLA-E bound to HLA class Ia signal sequence peptides with lower affinity than NKG2A) (13–15). The relevance of HLA-E/NKG2C interactions has been well-demonstrated in driving adaptive NK cell expansions and more recently a highly specific recognition of certain HCMV-encoded HLA-E presented peptides was elegantly shown (16, 17). A rare UL40 peptide, identical to the HLA-E-binding peptide in the HLA-G signal sequence, was found to trigger optimal NK stimulation and to have functional consequences, influencing NK cell antibody dependent cellular cytotoxicity (ADCC) responses (17).

It remains unclear how the presence of this polymorphism and changes in the HLA-E ligandome during infection and inflammation affect NK cell phenotypic and functional diversity in heterogenous populations with HIV-1 infection and high levels of HCMV co-infection. To further explore this, in the current study we investigated whether the HLA-B -21 dimorphism leads to a NK cell functional dichotomy in an African cohort co-infected with HIV-1/HCMV.

# MATERIALS AND METHODS

## Study Subjects

Cross-sectional analysis was performed on peripheral blood mononuclear cells (PBMCs) cryopreserved from chronically HIV-1 infected HCMV seropositive females recruited into the center for HIV/AIDS Vaccine Immunology (CHAVI)001 study at clinical sites in Africa. The CHAVI001 study was approved by the Duke Medicine and National Institutes of Health Institutional Review Boards as well as the ethics boards of the local sites. The subjects used for the work in this paper were all from the "established" infection group (defined as having a positive HIV antibody test, two concordant rapid HIV tests or standard EIA, and a fully positive Western blot profile, i.e., being at Fiebig stage 6 of infection, at the time of recruitment) of the CHAVI 001 study. They were all recruited at study sites in Africa: Blantyre and Lilongwe, Malawi; Durban and Johannesburg, South Africa; and Moshi, Tanzania. Exclusion criteria included the current use of antiretroviral treatment and any condition that, in the opinion of the Investigator of Record, would make participation in the study unsafe, complicate interpretation of study outcome data, or otherwise interfere with achieving the study objectives. All study participants gave written informed consent and were hepatitis C virus antibody negative and hepatitis B surface antigen (HBsAg) antibody negative. Human cytomegalovirus

(HCMV) infection status was determined by HCMV IgG enzyme-linked immunosorbent assay (ELISA) (BioKit) on stored plasma samples. HLA class I genotyping of the study donors to 2-digit allele resolution was performed by ProImmune (Oxford, UK) by PCR analysis of DNA extracted from donor PBMC. HLA-A expression model estimates (z-score) were inferred as previously described (11). The subject characteristics, HLA class I genotypes and distribution of HLA-B -21M and -21T among HLA-B groups are summarized in **Supplementary Table S1**.

## Monoclonal Antibodies and Flow Cytometry Analysis

For flow cytometric analysis, cryopreserved PBMC were thawed, washed in phosphate-buffered saline (PBS), and surface stained at 4°C for 20 min with saturating concentrations of different combinations of the following antibodies (**Supplementary Table S2**) in the presence of fixable Live/Dead stain (Invitrogen): CD14 BV510, CD19 BV510, CD56 PE Dazzle or CD56 BV605, CD3 BV650, CD16 PERCP, or CD16 BV711, HLA-E PE (3D12) (Biolegend), CD4-eFluor 780, CD8 Alexa700 (eBioscience), HLA-C PE (DT-9) (BD Biosciences), NKG2A Pe-Cy7, KIR2DL2 APC CD158b1/b2.j APC (Beckman Coulter), NKG2C PE or NKG2C Alexa 700, KIR2DL1/2DS5 APC IgG1 [CD158a], KIR3DL2 APC (R&D systems), CD57 BV421 or CD57 FITC (BD Biosciences), KIR3DL1 APC [CD158e1] (Miltenyi). For the detection of intracellular antigens cells were fixed, permeabilized and stained for IFN-γ BV421 (BD Biosciences) and FcεRI-γ-FITC (Millipore). The antibody against PLZF PE-CF594 (BD Biosciences) was used for intranuclear antigen detection utilizing the Foxp3 intranuclear staining buffer kit (eBioscience) according to the manufacturer's instructions. Samples were acquired on a BD Fortessa X20 using BD FACSDiva8.0 (BD Bioscience). Data were analyzed using FlowJo 10 (TreeStar) and stochastic neighbor embedding (SNE) analysis and FlowSOM analysis was performed on NK cells using the mrc.cytobank.org platform, utilizing the following parameters: CD16, KIRS, NKG2A, NKG2C, PLZF, Siglec-7, NKG2A, CD57, and FcεRI-γ. The FCS file concatenation tool was used for concatenating multiple FCS files into a single FCS file prior to uploading the files to Cytobank (Beckman Coulter).

## Functional ADCC Assay

For analysis of NK cell mediated ADCC responses, RAJI cells (1–2 × $10^6$ cells/ml) were coated with anti-CD20 or murine immunoglobulin G (IgG) (InvivoGen) at 2.5 µg/ml for 30 min. Subsequently RAJI cells were washed and then mixed with PBMCs in a V bottom 96-well plate at a 10:1 E:T ratio and incubated for 6 h 37C in the presence of CD107a-APC-H7 antibody (BD Biosciences, Cowley, U.K.). GolgiStop (containing Monensin, 1/1,500 concentration, BD Biosciences) and GolgiPlug (containing brefeldin A, 1/1,000 final concentration, BD Biosciences) were added for the last 5 h of culture. Following incubation cells were washed and stained for extracellular receptors prior to permeabilization and intracellular staining for IFN-γ. Boolean gating analysis was used to analyse CD107a and IFN-γ production in CD56$^{dim}$ NK cell subpopulations expressing CD57, NKG2A, and KIRs and combinations thereof.

## Soluble HLA-G Measurement

Soluble HLA-G1/G5 was measured in plasma by ELISA using a BioVendor-EXBIO kit according to the manufacturer's instructions.

## Data Analysis

Prism 7 (GraphPad Software) was used for all statistical analysis as follows: the Mann-Whitney *U*-test or Student's *t*-test were used for single comparisons of independent groups, the Wilcoxon-test was used to compare two paired groups and the Kruskal-Wallis with Dunn's multiple comparison test was used to compare three unpaired sample groups. The non-parametric Spearman test was used for correlation analysis. SPICE analysis was performed in SPICE version 6. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$.

# RESULTS

## Haplotypes Combining HLA-C2 and -21M HLA-B Are Common in African Populations and the HLA-B -21M Dimorphism Does Not Significantly Impact on Surface HLA-E Expression

To explore the effects of the HLA-B dimorphism in a non-Caucasian population, we initially analyzed HLA haplotypes and examined the segregation of HLA-C allotypes and -21 HLA-B alleles in a cohort of viraemic age-matched HIV-1 infected HCMV-seropositive African females, representing the three key -21 HLA-B genotypes: -21M/M homozygotes, -21M/T heterozygotes, and -21T/T homozygotes (**Figure 1A**). There were no significant differences in the HIV-1 viral load levels between the three groups (**Supplementary Table S1**). In contrast to Eurasian populations, which have an effective exclusion of -21M HLA-B from haplotypes encoding HLA-C2, this segregation was not evident in this cohort (**Figure 1A**), in keeping with the presence of African specific alleles, B*42:01–C*17:01 and B*81:01–C*18:01 in the M/M group (**Supplementary Table S1**). Such haplotypes combining HLA-C2 with -21M HLA-B provide both a C2 allele, a stronger KIR ligand than C1, and an HLA-E ligand for NKG2A. HLA-B -21M alleles did not encode HLA-B Bw4 in our cohort, in line with data derived from larger population analysis (6) but interestingly, a high proportion of the subjects with -21T HLA-B alleles (nine out of 13 subjects) also did not encode HLA-B Bw4, which functions as a KIR ligand (**Supplementary Table S1**). The subsets of HLA haplotypes in the study groups defined by the presence of -21M HLA-B in various combination with HLA-C1 and C2 could therefore result in the availability of KIR ligands differentially supplying HLA-E-binding peptides to form NKG2A ligands being distinct from that in Caucasian populations, with consequences for NK cell education.

To investigate the effects of the HLA-B -21 dimorphism on surface expression of HLA-E we examined the expression

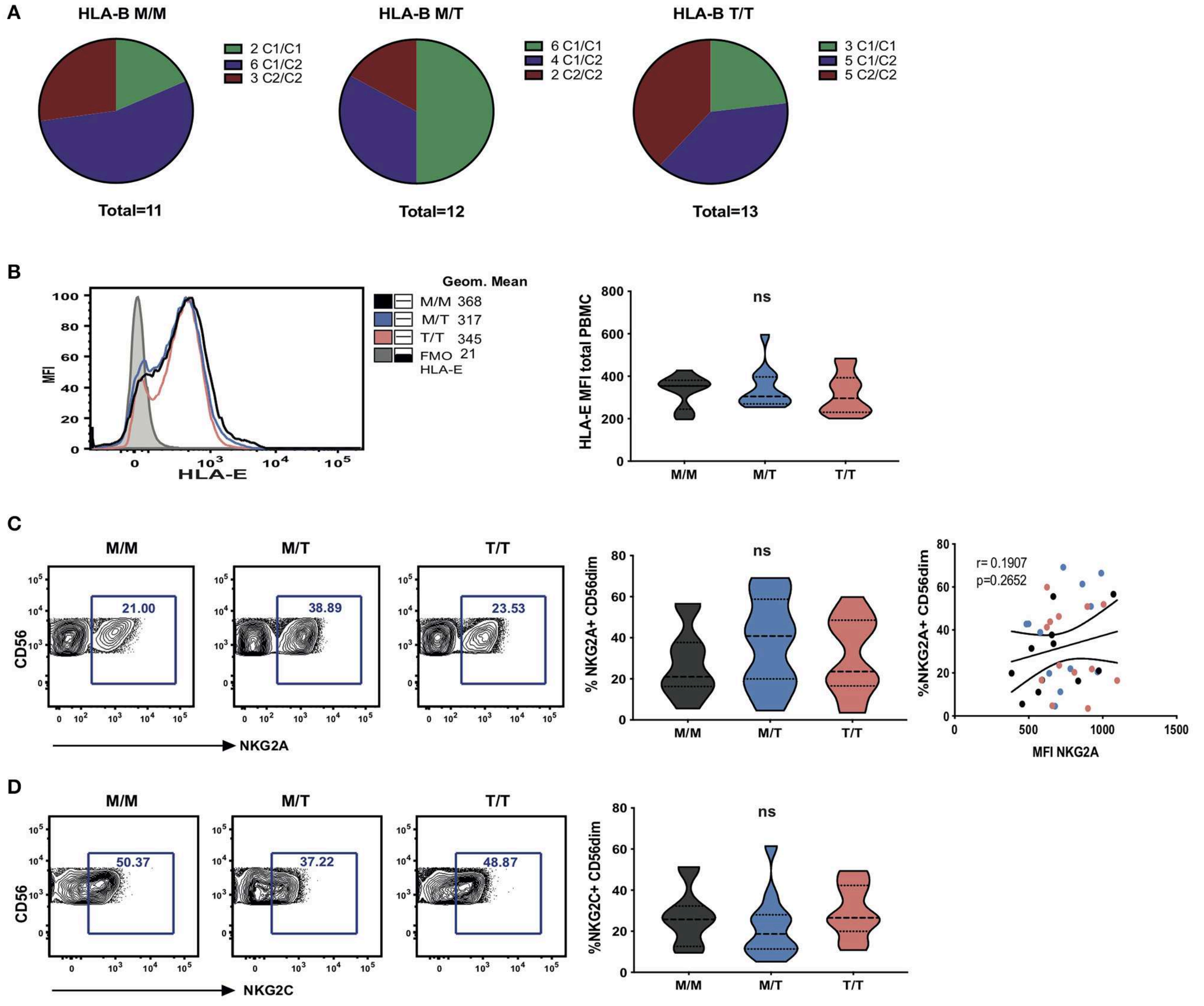


**FIGURE 1 |** Dimorphism at position -21 HLA-B does not significantly modulate HLA-E and NKG2A expression. **(A)** HLA haplotypes encoding HLA-C1 and C2 within groups of -21 HLA-B genotype M/M homozygous, -21M/T heterozygous and -21 T/T homozygous subjects from the study cohort. **(B)** Representative histograms showing HLA-E expression on total PBMC between groups as well as fluorescence minus one (FMO) control staining (left); and comparison of cell-surface HLA-E expression (geometric mean fluorescence intensity (MFI) of staining with HLA-E-specific antibody 3D12) on total PBMC between groups (right). Data are displayed as violin plots; the group median and interquartile range are indicated. **(C)** Proportion of $CD56^{dim}$ NK cells (i.e within live CD56+CD3-CD19-CD14-CD4- PBMC) expressing NKG2A between groups and comparison of the percentage of NKG2A+ $CD56^{dim}$ with their level of surface expression (MFI) in M/M (black dots), M/T (blue dots) and T/T (pink dots) subjects. **(D)** Violin plots of the frequency of NKG2C expressing $CD56^{dim}$ NK cells in the donor groups. Group median and interquartile range are indicated. All donors were HCMV positive. ns: non-significant.

of HLA-E on peripheral blood mononuclear cells (PBMCs) in subject groups distinguished on the basis of the amino acid encoded (M/M, M/T, and T/T). Despite median levels of total cellular HLA-E expression on PBMCs being higher in M/M individuals, no significant difference was observed in surface HLA-E expression on PBMC between the groups (**Figure 1B**). Further analysis of surface HLA-E expression on CD3+, CD4+, and CD8+ T cells and CD3- cells did not show any significant differences between the groups (**Supplementary Figure S1A**). Moreover, no correlation was detected between -21M copy number and the proportion of NK cells expressing either NKG2A (**Figure 1C**) or its activating counterpart NKG2C (**Figure 1D**). These observations contrast findings in HCMV seronegative Eurasian individuals where HLA haplotypes defined by -21M HLA-B were associated with increased surface expression of HLA-E and a decreased frequency of NK cells expressing NKG2A (6).

Furthermore, no relationship was detected between the surface levels of HLA-E expression and HLA-A imputed expression level (z score) in subject groups distinguished on the basis of the presence of HLA-B -21M in this cohort (**Supplementary Figure S1B**). The effect of HLA-A expression on HIV-1 viraemia was also not evident in individuals with HLA-B -21M/M, in keeping with a reported less prominent effect in African/African-Americans relative to Caucasians, conceivably as a consequence of the distinct HLA haplotypes and frequencies present in individuals of African descent (**Supplementary Figure S1C**) (11).

## Surface Abundance of HLA-C Does Not Correlate With -21 HLA-B

The amount of cell-surface HLA-C expression has been previously reported to vary with -21 HLA-B type, with M/M HCMV seronegative European donors displaying low levels, as a result of their relatively restricted HLA-C diversity and genotypes dominated by HLA-C*07, which is subject to microRNA-148a (miR-148a) mediated downregulation (18). Whilst in predominantly Eurasian populations haplotypes combining -21M HLA-B and C2 are rare, the African haplotypes present in the M/M group combine specific HLA-B and C2 alleles, i.e., B*42:01–C*17:01 and B*81:01–C*18:01. Despite some variation in the levels of surface expression of HLA-C (assessed by staining with the HLA-C and HLA-E-reactive antibody DT9) (19), especially in T/T donors, there was no overall difference in the mean levels of expression between the study groups, and there was no correlation between cell surface abundance of HLA-C and -21 HLA-B genotype on total PBMC (**Figures 2A,B**) and lymphocyte subsets (data not shown). Equally, we observed no obvious clustering according to HLA-C1 and C2 types and only a single M/T donor was homozygous for HLA-C*07, an allele that is highly represented in M/M individuals of European origin as previously shown (**Figures 2C,D**) (6).

## KIR Education and Differentiation Predominate in HIV-1/HCMV Seropositive Subjects of African Descent Irrespective of -21 HLA-B Dimorphism

The effect of viraemic HIV-1 infection on driving alterations in the NK cell subset distribution is well-described (15, 20). In keeping with this we confirmed the presence of the aberrant CD56$^{neg}$CD16+ NK cell subset in our cohort; however, no significant differences in the frequencies of the CD56$^{bright}$, CD56$^{dim}$, and CD56$^{neg}$ NK cell subsets were observed between groups of M/M, M/T, and T/T individuals (**Supplementary Figure S2A** gating strategy and **Supplementary Figures S2B–D**). Notably chronic HIV-1/HCMV co-infection also leads to an accentuated differentiation within the CD56$^{dim}$ subset with the emergence of a CD57+NKG2C+KIR+NKG2A- signature and expansion of adaptive NK cell subsets (15, 21, 22). We therefore investigated the phenotypic diversity of CD56$^{dim}$ NK cell subset to delineate the fingerprint of HCMV co-infection in the three study groups in relation to the presence of -21 HLA-B dimorphism.

In the cohort as a whole, the acquisition of inhibitory KIRs on CD56$^{dim}$ NK cells was paralleled by a loss of NKG2A expression ($r = -0.4165$, $p = 0.0143$) (23). A tight positive correlation between NKG2C and KIR expression was further noted, in line with the expansion of self-specific KIRs in the context of HCMV infection/re-activation ($r = 0.5960$, $p = 0.0002$) (24–26). The level of expression of KIRs (cocktail of antibodies against KIR2DL1/S5, KIR2DL2/L3/S2, KIR3DL2, and KIR3DL1) on CD56$^{dim}$ NK cells did not differ between the three study groups (**Figures 3A,B**). Levels of CD57 expression were also comparable between the three groups, suggesting the presence of NK cells at different differentiation stages (**Figures 3A,B**). Examination of additional markers such as the key signaling molecule FcεRI-γ and the transcription factor promyelocytic leukemia zinc (PLZF), the absence of which characterizes adaptive NK cell subsets, showed a broader range of expression with a trend for a lower median level of expression in T/T subjects, which did not reach statistical significance (**Figures 3A,B**).

Boolean gating analysis was performed next to examine the proportion of NKG2A or KIR-educated CD56$^{dim}$ NK cells that were more highly differentiated (assessed on the basis of expression of CD57) (**Figure 3C** representative viSNE analysis showing clusters of NKG2A and KIRs co-expressing CD57 and SPICE analysis). The proportion of KIR-NKG2A+ CD56$^{dim}$ NK cells, educated via the inhibitory NKG2A receptor, was 21.95% ± 4.432 (mean ± *SEM*) in M/M, 29.09% ± 5.315 in M/T, and 26.26% ± 5.432 in T/T donors. The extent to which these NKG2A-educated cells were differentiated (according to the expression of CD57) did not vary with the -21M copy number and represented a small fraction (10.65% ± 2.440, mean ± *SEM* in M/M donors, 12.89% ± 3.262 in M/T, and 14.12% ± 3.508 in T/T donors). KIR+NKG2A- NK cells, which can only be educated via KIRs, represented a larger fraction of CD56$^{dim}$ NK cells in all groups (45.83% ± 6.159, 36.57% ± 6.847, and 49.48% ± 5.767, in M/M, M/T, and T/T subjects, respectively). The proportion of KIR-educated NK cells that were differentiated (CD57+) was higher than that observed in the NKG2A-educated fraction, comprising in M/M donors 40.43% ± 5.204 in M/M donors and 40.82 ± 5.339 in T/T donors, and trending to be somewhat lower in M/T subjects (28.59 ± 5.629) (**Figure 3C** pie charts). These results are in keeping with loss of NKG2A expression with increasing NK cell differentiation in HIV infection and contrast findings of a dominant effect of the -21M HLA-B dimorphism on increasing the differentiated subpopulation of the educated KIR-NKG2A+ NK cells in Caucasian HCMV seronegative donors (6).

Whereas, all donors exhibited comparable levels of CD57 expression on CD56$^{dim}$ NK cells, the activating receptor CD16 was expressed by a higher proportion of differentiated CD57+ CD56$^{dim}$ NK cells in T/T donors compared to M/M and M/T donors ($p = 0.02$ and $p = 0.01$, respectively; **Figure 3D**). As expected the CD57+ subset of NK cells in all groups was enriched for additional adaptive features such as lower levels of PLZF and FcεRI-γ and enriched for KIR and NKG2C compared to the CD57 negative fraction of NK cells, as previously described (15). In T/T donors the CD57+ NK cell subset trended to have higher mean levels of expression of NKG2C and lower mean levels of expression of PLZF and FcεRI-γ compared to the CD57+ NK cells in M/M and M/T donors although this did not reach statistical significance (**Supplementary Figure S3A**). Analysis with self-organizing maps (FlowSOM) did not demonstrate

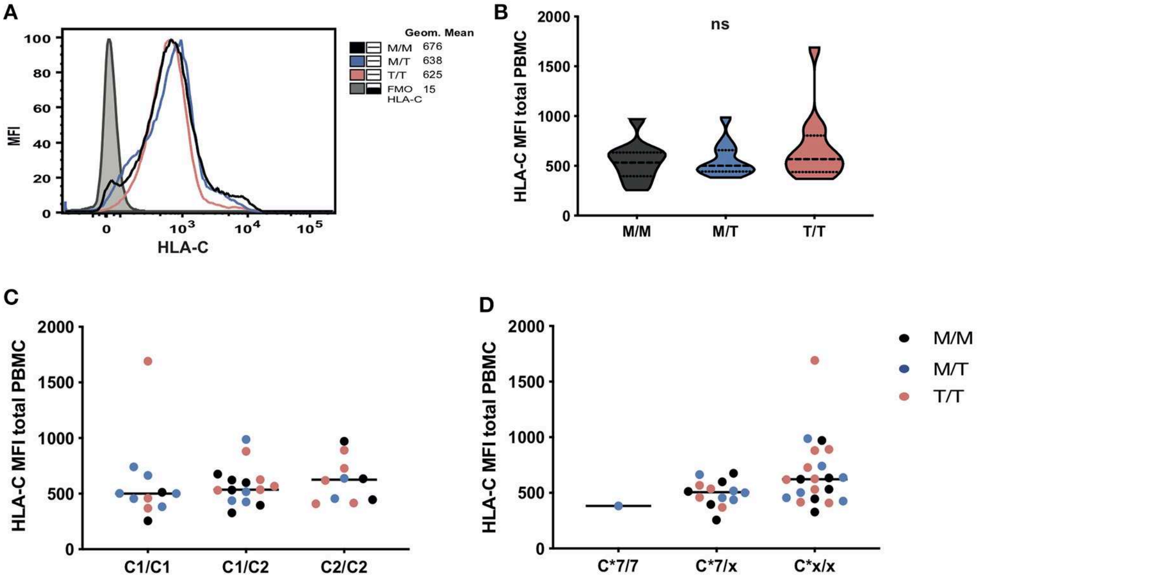


**FIGURE 2 |** No significant differences in HLA-C expression levels in subjects grouped according to -21 HLA-B dimorphism. **(A)** Representative histograms showing HLA-C expression (geometric mean fluorescence intensity (MFI) on total PBMC between groups as well as fluorescence minus one (FMO) control staining). Levels of surface expression of HLA-C (MFI of staining with antibody DT-9) on total PBMC in donors grouped: **(B)** by HLA-B -21 variant (data are shown as violin plots, and group median and quartiles are indicated); **(C)** according to HLA-C1 and C2 epitopes and **(D)** by the presence and absence of (x) of HLA-C*07. In panels **(C)** and **(D)**, M/M subjects are shown as black dots, M/T as blue dots and T/T as pink dots.

any prominent clustering differences depending on the HLA-B dimorphism (**Supplementary Figures S3B,C**).

## The Dominant Effect of -21M on NK Cell Function Is Lost in HIV-1/HCMV Donors

To further examine the influence of -21 HLA B dimorphism on NK cell education and associated NK cell function we utilized an antibody-coated target cell stimulation assay to measure ADCC. Following stimulation with Raji cells coated with anti-CD20, CD56$^{dim}$ NK cells were assessed for cytokine production by intracellular cytokine staining and degranulation, as measured by surface expression of CD107a. NK cells from T/T donors demonstrated a trend toward higher production of IFN-γ relative to those from M/M donors and higher IFN-γ production compared to M/T individuals (**Figure 4A**). A similar trend toward higher NK cell expression of CD107a was observed in T/T donors in relation to the M/M and M/T groups (**Figure 4B**). For both functional responses a range of IFN-γ production and CD107a expression was observed within each group that could not be attributed to their HLA-C haplotype (data not shown). Further analysis of the proportion of IFN-γ producing CD56$^{dim}$ NK cells that were differentiated (according to CD57 expression) and either educated via NKG2A or KIRs, showed a higher proportion of the cytokine producing cells being comprised of CD57+NKG2A-KIR+ NK cells than CD57+NKG2A+KIR- cells in M/M and T/T donors ($p = 0.006$ and $p = 0.005$, respectively) but this did not reach statistical significance for the M/T group. These differences are reflected in all three pie charts (**Figure 4C**), where the subset of IFN-γ-producing cells with a CD57+NKG2A-KIR+ phenotype is of similar size in M/M, T/T, and slightly smaller in M/T donors, in keeping with lower frequencies of differentiated KIR educated NK cells in this group. These data suggest the dominant effect of KIR mediated education for IFN-γ producing NK cells irrespective of -21 HLA-B dimorphism. This is in contrast to the gene dosage effect of -21 HLA-B dimorphism on KIR educated NK cells in HCMV seronegative HIV negative Europeans, where two copies of -21T results in a 2.6-fold increase in the number of KIR+NKG2A-NK cells producing IFN-γ compared to that in M/M subjects (11). A similar KIR predominant effect was observed for CD107a production in M/M and T/T subjects in our cohort, whereas in M/T donors the effects of the educating KIRs were less distinct (data not shown).

In addition to the effect of education, the variability in the ADCC functional responses of CD56$^{dim}$ NK cells, in particular IFN-γ production, could relate to cellular expression of CD16. *Ex vivo*, the proportion of NK cells expressing CD16 correlated with IFN-γ production ($r = 0.5609$, $p = 0.0007$), suggesting that shedding of CD16 reported in progressive HIV-1 infection, could contribute to the reduction of NK cell ADCC function in our study cohort (**Figure 4D**) (27). Notably adaptive NKG2C+ NK cell subpopulations that arise in response to HCMV infection and expand during HIV-1 infection are imbued with enhanced ADCC capacity, in particular production of IFN-γ following CD16 ligation, reflecting epigenetic modifications and enhanced downstream signaling through CD3z homodimers in the absence of FcεRI-γ (15, 28). We therefore assessed whether the size of adaptive NKG2C+ NK cell populations could account for the variability in IFN-γ production noted between and within the three study groups. In the cohort as a

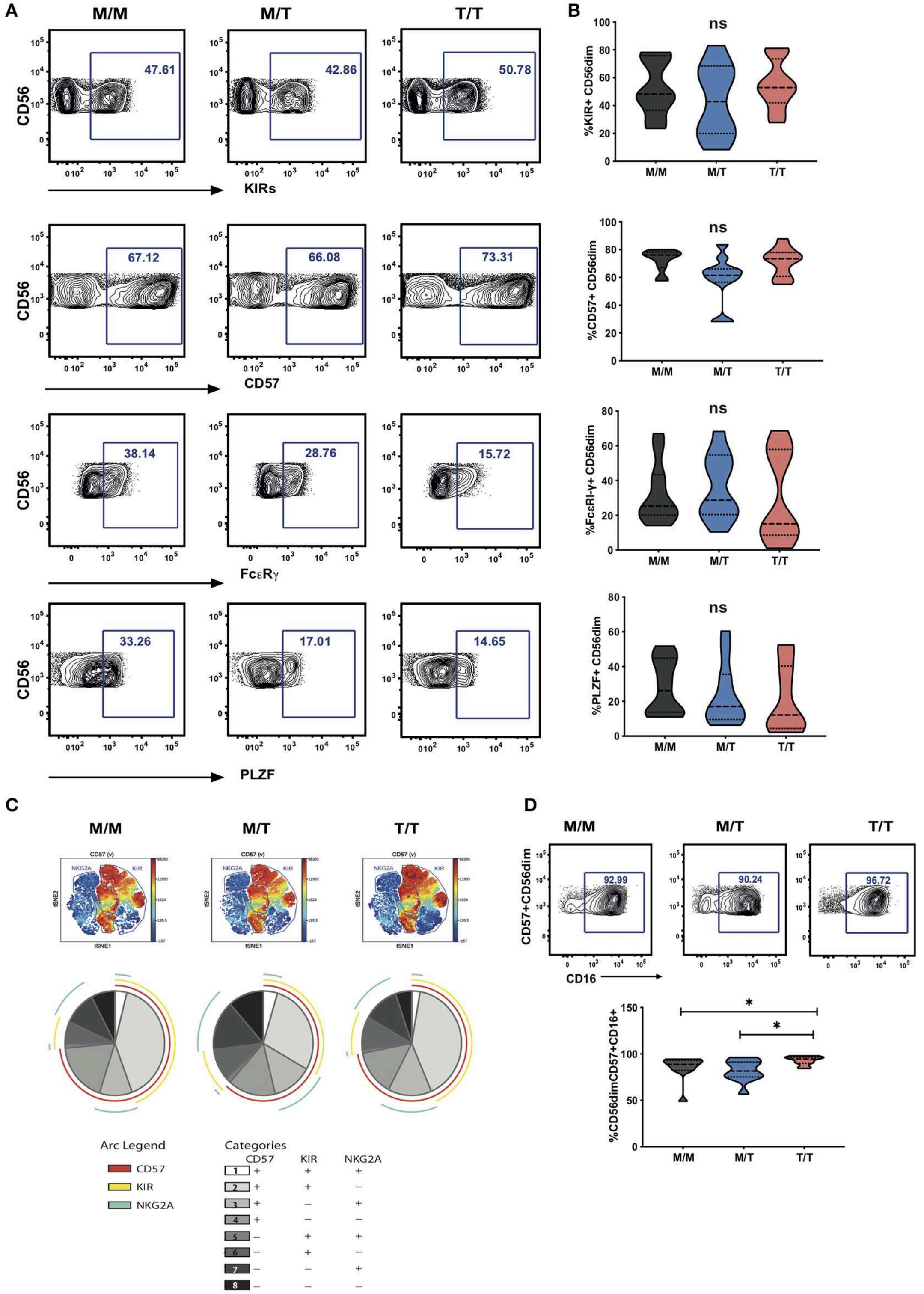


**FIGURE 3 |** Lack of a dominant effect of -21M HLA-B on the extent of NKG2A driven NK cell differentiation. **(A)** Representative contour plots depicting the gating strategy for the expression of KIRs, CD57, FcεRI-γ and PLZF in CD56$^{dim}$ NK cells from each study group. **(B)** Summary violin plots of the proportion of CD56$^{dim}$ NK cells expressing KIRs, CD57, FcεRI-γ and PLZF between donor groups. **(C)** ViSNE analysis of multiparametric flow data was performed on CD56$^{dim}$ NK cells from the compiled M/M, M/T and T/T donors showing expression of CD57 and gated KIR and NKG2A clusters. Each point on the VisNE map represents a single cell and color depicts intensity of protein expression. SPICE analysis pie charts of CD57, KIR and NKG2A expression for each group. The pie slices represent the proportion of CD56$^{dim}$ NK cells expressing different receptor combinations, and the pie arcs depict expression of individual receptors, as detailed in the key. **(D)** Representative contour plots and summary violin plots showing the proportion of CD57+CD56$^{dim}$ NK cells expressing CD16 between the donor groups. *$p < 0.05$.

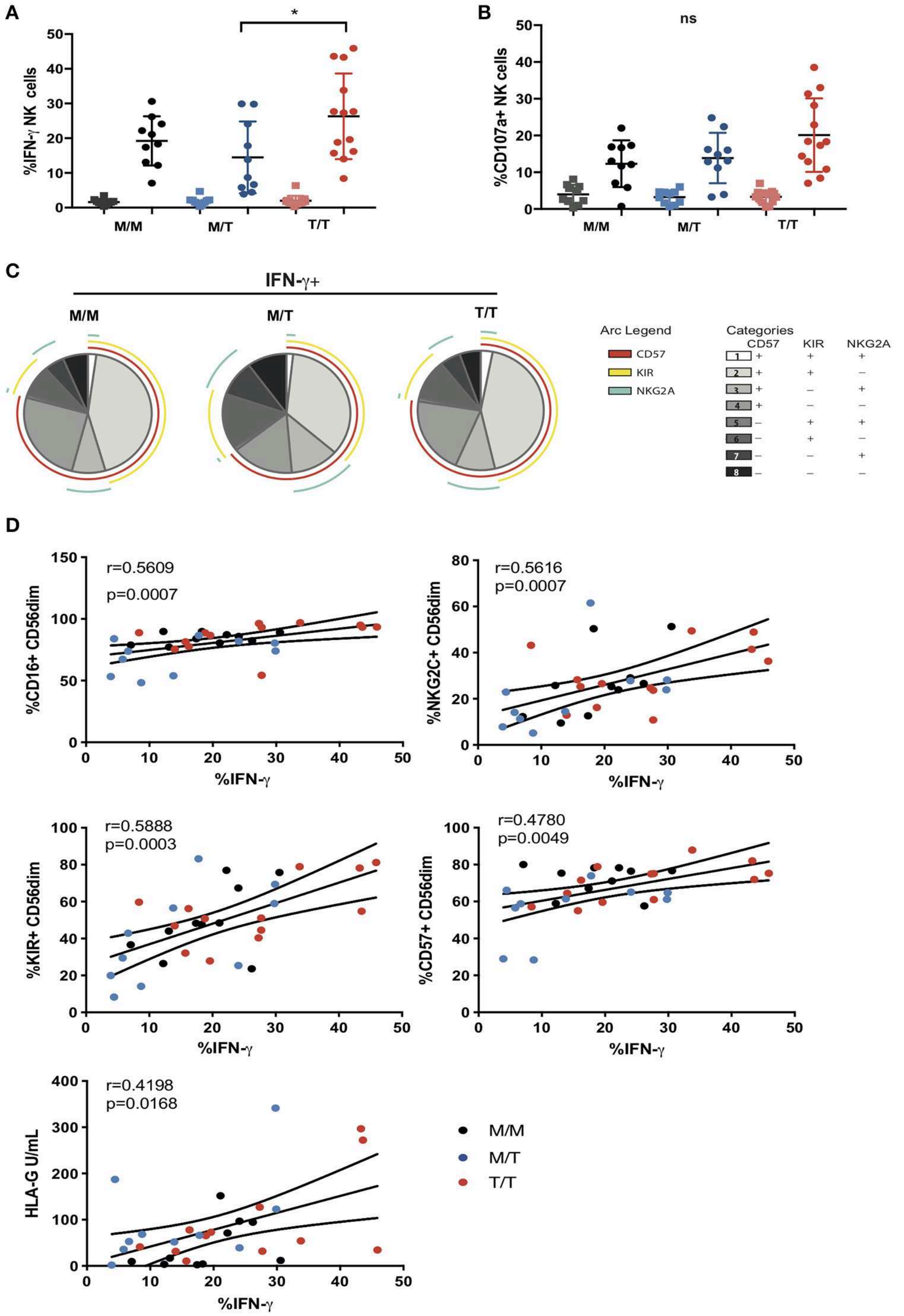


**FIGURE 4 |** Variable influence of -21 HLA-B dimorphism on ADCC responses. **(A)** IFN-γ and **(B)** CD107a expression by CD56$^{dim}$ NK cells following co-culture with RAJI cells coated with anti-CD20 (filled circles) or murine IgG (filled squares) in M/M ($n$ = 10), M/T ($n$ = 10) and T/T donors ($n$ = 13) with available PBMC. **(C)** SPICE pie charts for each group. The pie slices correspond to the proportion of IFN-γ producing cells that express different receptor combinations, and the pie arcs depict individual expression of CD57, KIRs, and NKG2A, as detailed in the key. **(D)** Correlations between IFN-γ production by CD56$^{dim}$ NK cells and expression of CD16, CD57, KIRs, NKG2C, or soluble HLA-G levels. The non-parametric Spearman test was used for correlation analysis.

whole, IFN-γ production correlated strongly ($r = 0.5616$, $p = 0.0007$) with NKG2C expression, suggesting that the presence of adaptive subpopulations, enriched within differentiated CD57+ and KIR+ NK cells (which also correlate with IFN-γ production), could modulate NK cell functional capacity to antibody coated targets (**Figure 4D**).

Recently it was demonstrated that HCMV-derived peptides presented by HLA-E, in particular the rare UL40 peptide VMAPRTLFL which is identical to the HLA-G leader peptide, fine tune the ADCC response of NK cells via NKG2C recognition (17). Both membrane bound and soluble levels of HLA-G are reported to be increased in untreated HIV-1 infection and during HCMV infection and have been shown to correlate with blood IFN-$\gamma$ concentrations and could therefore represent a source of HLA-E peptides in T/T individuals (29–31). Although we did not detect any significant differences in the soluble plasma HLA-G concentration between the study groups, HLA-G levels showed a weak association with IFN-$\gamma$ production suggesting that an environment potentially displaying altered HLA-E peptide ligands recognized by adaptive NKG2C expressing NK cells may induce differential cellular responses (**Figure 4D**).

## DISCUSSION

HLA-E acts as powerful modulator of the immune response, serving as a ligand for NKG2 receptors that provide a functionally complementary axis to the polymorphic KIR system for control of innate lymphocyte subsets. HLA-E binds signal peptides derived from the leader sequence of HLA-A, B, C, and G proteins in order to achieve stable expression at the cell surface (32). -21M, the residue present in all HLA-A and -C and a minority of -B allotypes, facilitates folding and expression of HLA-E by providing a strong anchor residue in contrast to -21T, the residue present in the majority of HLA-B allotypes. This genetic segregation depending on HLA-B dimorphism leads to a binary form of NK cell education and functional responsiveness in HCMV seronegative donors of European origin by either supplying NKG2A or KIR ligands (6). A similar effect was not seen in an African cohort with HIV-1/HCMV co-infection, where genetic and environmental factors could influence the NK cell repertoire and effector function. The presence of African specific alleles, together with alterations in the HLA-E peptide repertoire due to the availability of peptides derived from other cellular and viral sources that could arise during HIV-1/HCMV coinfection, trigger the expansion of adaptive NK cells expressing the activating receptor NKG2C with subsequent functional consequences. The lack of -21M expression could thus become redundant in HCMV seropositive individuals where UL40 or HLA-G derived peptides may stabilize the expression of HLA-E and fine tune NK cell activation and antibody driven adaptive responses.

In Eurasian populations the reported LD between HLA-B -21M and HLA-B Bw6/HLA-C1 limits the supply of KIR ligands and favors NKG2A mediated NK cell education (6). However, the genetic segregation between HLA-C1 and -21M HLA-B was not evident in this study group, where the presence of HLA-C2, a stronger KIR ligand than C1, resulted in the presence of both KIR and HLA-E ligands for NKG2 receptors in M/M donors. In addition, the more common African haplotypes combining -21M and HLA-C2 involve African specific HLA-C allotypes that have leader sequences that poorly promote HLA-E expression, further limiting the supply of HLA-E ligands for interaction with NKG2 receptors on NK cells. These genetic effects could partly explain the lack of association between -21M copy number and surface HLA-E expression in our cohort. Another possible genetic factor that may have influenced the levels of HLA-E expression in our cohort is the dimorphism at position 107 of HLA-E, which distinguishes two most common alleles, HLA-E*01:01 (position 107 arginine, R) and HLA-E*01:03 (position 107 glycine, G), the former of which is reported to be expressed at lower levels than the latter (33) although this has not been seen in all studies (9). Nonetheless, although HLA-E genotyping was not performed in our cohort, as the two main HLA-E alleles occur in roughly equal frequencies in different ethnic groups and are maintained in diverse HLA haplotypes by stabilizing selection (34), allele frequencies would not have been expected to differ significantly between our study groups.

In addition to genetic differences between our cohort and those studied previously, the presence of chronic HIV-1 and HCMV co-infection in our study subjects may also have contributed to the lack of significant difference in surface HLA-E expression between study groups. HLA-E surface levels serve as an important sensor of HLA class I expression and are sensitive to perturbations in the biosynthesis of most polymorphic class I allotypes as well as the class Ib molecule HLA-G imparted by viral infections or stress. Of note whilst HIV-1 Nef causes down-regulation of HLA-A, B and Vpu mediates reduction of HLA-C, these viral accessory proteins mediate their effects post-translationally and should not affect the supply of HLA class I signal peptides; and HCMV maintains/stabilizes HLA-E expression (35–38). However, the presence of specific HCMV UL40 variants and/or HLA-G levels may be altering the supply of HLA-E binding peptides in our cohort.

As well as observing no impact of the HLA-B -21 dimorphism on the level of expression of HLA-E we also did not detect a correlation between -21M copy number and NKG2A expression in our cohort. During NK cell development and education, the acquisition of self-reactive KIRs leads to progressive downregulation and decreased surface expression of NKG2A (23). This process is accelerated during HIV-1 infection/HCMV co-infection and further underlined by the expansion of differentiated CD57+ NKG2C+ NK cell subsets enriched for KIRs for self HLA-C1 and/or C2 allotypes, which explains the lack of correlation between -21M HLA-B and better NKG2A+KIR- educated NK cells in this cohort. Due to limitations in sample availability we were not able to type the KIR genes nor perform staining for individual KIRs in our study subjects. Specific KIR alleles are reported to differ in their strength of signaling, with associated effects on NK cell education/ADCC responses, which could further explain some of the inter-donor variability observed in this study. A recent study has further highlighted the critical role of KIR polymorphism influencing responses to HCMV, where in particular the interaction between KIR2DL1 and HLA-C2 ligands drives large and stable expansions of adaptive NKG2C+ NK cells

(26). It would therefore be of interest to determine the effect of KIR polymorphism in modulating the size of the adaptive NK cell pool in larger HIV-1/HCMV co-infected cohorts.

There is increased appreciation that peptides presented via HLA-E during conditions of stress and viral infections influence the activation of NK cells, driving expansion of adaptive NKG2C+ NK cells and subsequent enhancement of ADCC responses. In keeping with this, we observed a range of ADCC responses in our cohort that correlated with NKG2C expression. Furthermore, UL40 in HCMV encodes peptides that mimic MHC class I signal sequences and share a conserved methionine (M) anchor residue at peptide amino-acid residue 2), which correspond to amino acid -21 of the classical HLA class I leader sequence (16, 17). Hence HCMV infection provides peptides that may substitute for host HLA-I-derived HLA-E stabilizing non-americ peptides in T/T donors. Interestingly UL40 polymorphisms and the strength of interactions between HLA-E presented peptides and NKG2C controls the activation of adaptive NK cells. Of note, a gradient in NKG2C+ NK cell effector function has been reported depending on the potency of recognition of HCMV peptides (VMAPRTLFL > VMAPRTLIL > VMAPRTLVL) (17). The VMAPRTLFL UL40 derived peptide mimics the signal peptide of HLA-G, the expression of which is upregulated during inflammation, HCMV infection and HIV-1 infection, and specifically enhances antibody-driven adaptive NK cell responses as recently described (17). Regardless of the peptide source, it is tempting to speculate that alterations in the HLA-E ligandome surveyed by the NKG2C receptor contribute differentially to the accumulation, differentiation and effector functions of adaptive NKG2C+ NK cells during infection. Whether HIV-1 peptides could further exploit the HLA-E/NKG2 axis as recently suggested (39) requires further evaluation.

The role of adaptive NK cells in influencing the rate of HIV-1 acquisition and levels of viral control during established infection remains poorly defined, but offers an alternative explanation for previous epidemiological observations, that needs to be formally addressed with a combination of population and functional studies. Assessment of the overall impact of HLA-B dimorphism on the acquisition of or control of HIV-1 infection will need to take into account a number of effects in addition to the contribution of HLA/peptide complex availability and its impact on the NKG2 pathway. These include effects on CD8 T cell responses and interactions between the Bw4/Bw6 epitope and the KIR3DL1/3DS1 pathway, which will necessitate study of much larger cohorts.

In summary, we posit that in addition to differences in the genetic background, chronic HIV-1 infection with frequent reactivations of HCMV affects the pool of peptides presented by HLA-E and surface levels of HLA-E providing a more diverse range of ligands for CD94/NKG2 NK cells. The strength of these interactions and presence of inflammatory stimuli shape the NK cell pool and functional activity, blurring the dichotomous effect of -21 HLA-B on NK cell function seen in Eurasian HCMV seronegative donors. Future larger studies aimed at dissecting the effect of different HLA-E/peptide ligands on adaptive NK cells in relation to -21 HLA-B polymorphism, during disease are required, in order to facilitate realization of the translational potential of specific NK cell subpopulations and exploit the NKG2C/HLA-E axis to enhance NK cell functionality.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## ETHICS STATEMENT

The CHAVI001 study was approved by the Duke Medicine and National Institutes of Health Institutional Review Boards as well as the ethics boards of the local sites. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

EC: performed experiments, contributed to study design, acquisition of data, analysis, and drafting of the manuscript. AO and IP-P: performed experiments and contributed to acquisition of data. MC, BH, and PB: contributed to study design, data interpretation, and critical editing of the manuscript. DP: conception and design of study, data analysis and interpretation, critical revision of the manuscript and study supervision.

## FUNDING

This work was supported by MRC grants MR/M008614 (DP) and MR/K012037 (PB), an AOP grant awarded to DP and NIH, NIAID, DAIDS UM1 grants AI67854 (CHAVI), AI00645 (Duke CHAVI-ID) and AI144371 (BH and PB) and R01 award AI147778 (PB and DP). PB was a Jenner Institute Investigator.

## ACKNOWLEDGMENTS

This manuscript has been released as a preprint at bioRxiv (40).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00156/full#supplementary-material

## REFERENCES

1. Waggoner SN, Reighard SD, Gyurova IE, Cranert SA, Mahl SE, Karmele EP, et al. Roles of natural killer cells in antiviral immunity. *Curr Opin Virol*. (2016) 16:15–23. doi: 10.1016/j.coviro.2015.10.008
2. O'Sullivan TE, Sun JC, Lanier LL. Natural Killer cell memory. *Immunity*. (2015) 43:634–45. doi: 10.1016/j.immuni.2015.09.013

3. Scully E, Alter G. NK cells in HIV disease. *Curr HIV/AIDS Rep*. (2016) 13:85–94. doi: 10.1007/s11904-016-0310-3
4. Boudreau JE, Hsu KC. Natural Killer cell education in human health and disease. *Curr Opin Immunol*. (2018) 50:102–11. doi: 10.1016/j.coi.2017.11.003
5. Braud VM, Allan DS, O'Callaghan CA, Soderstrom K, D'Andrea A, Ogg GS, et al. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature*. (1998) 391:795–9. doi: 10.1038/35869
6. Horowitz A, Djaoud Z, Nemat-Gorgani N, Blokhuis J, Hilton HG, Beziat V, et al. Class I HLA haplotypes form two schools that educate NK cells in different ways. *Sci Immunol*. (2016) 1:eaag1672. doi: 10.1126/sciimmunol.aag1672
7. Vales-Gomez M, Reyburn HT, Erskine RA, Lopez-Botet M, Strominger JL. Kinetics and peptide dependency of the binding of the inhibitory NK receptor CD94/NKG2-A and the activating receptor CD94/NKG2-C to HLA-E. *EMBO J*. (1999) 18:4250–60. doi: 10.1093/emboj/18.15.4250
8. Merino AM, Song W, He D, Mulenga J, Allen S, Hunter E, et al. HLA-B signal peptide polymorphism influences the rate of HIV-1 acquisition but not viral load. *J Infect Dis*. (2012) 205:1797–805. doi: 10.1093/infdis/jis275
9. Merino AM, Sabbaj S, Easlick J, Goepfert P, Kaslow RA, Tang J. Dimorphic HLA-B signal peptides differentially influence HLA-E- and natural killer cell-mediated cytolysis of HIV-1-infected target cells. *Clin Exp Immunol*. (2013) 174:414–23. doi: 10.1111/cei.12187
10. Flores-Villanueva PO, Yunis EJ, Delgado JC, Vittinghoff E, Buchbinder S, Leung JY, et al. Control of HIV-1 viremia and protection from AIDS are associated with HLA-Bw4 homozygosity. *Proc Natl Acad Sci USA*. (2001) 98:5140–5. doi: 10.1073/pnas.071548198
11. Ramsuran V, Naranbhai V, Horowitz A, Qi Y, Martin MP, Yuki Y, et al. Elevated HLA-A expression impairs HIV control through inhibition of NKG2A-expressing cells. *Science*. (2018) 359:86–90. doi: 10.1126/science.aam8825
12. Gianella S, Massanella M, Wertheim JO, Smith DM. The sordid affair between human herpesvirus and HIV. *J Infect Dis*. (2015) 212:845–52. doi: 10.1093/infdis/jiv148
13. Guma M, Budt M, Saez A, Brckalo T, Hengel H, Angulo A, et al. Expansion of CD94/NKG2C+ NK cells in response to human cytomegalovirus-infected fibroblasts. *Blood*. (2006) 107:3624–31. doi: 10.1182/blood-2005-09-3682
14. Beziat V, Liu LL, Malmberg JA, Ivarsson MA, Sohlberg E, Bjorklund AT, et al. NK cell responses to cytomegalovirus infection lead to stable imprints in the human KIR repertoire and involve activating KIRs. *Blood*. (2013) 121:2678–88. doi: 10.1182/blood-2012-10-459545
15. Peppa D, Pedroza-Pacheco I, Pellegrino P, Williams I, Maini MK, Borrow P. Adaptive reconfiguration of Natural Killer cells in HIV-1 infection. *Front Immunol*. (2018) 9:474. doi: 10.3389/fimmu.2018.00474
16. Hammer Q, Ruckert T, Borst EM, Dunst J, Haubner A, Durek P, et al. Peptide-specific recognition of human cytomegalovirus strains controls adaptive Natural Killer cells. *Nat Immunol*. (2018) 19:453–63. doi: 10.1038/s41590-018-0082-6
17. Rolle A, Meyer M, Calderazzo S, Jager D, Momburg F. Distinct HLA-E peptide complexes modify antibody-driven effector functions of adaptive NK cells. *Cell Rep*. (2018) 24:1967–76e4. doi: 10.1016/j.celrep.2018.07.069
18. Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, et al. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*. (2011) 472:495–8. doi: 10.1038/nature09914
19. Corrah TW, Goonetilleke N, Kopycinski J, Deeks SG, Cohen MS, Borrow P, et al. Reappraisal of the relationship between the HIV-1-protective single-nucleotide polymorphism 35 kilobases upstream of the HLA-C gene and surface HLA-C expression. *J Virol*. (2011) 85:3367–74. doi: 10.1128/JVI.02276-10
20. Mavilio D, Lombardo G, Benjamin J, Kim D, Follman D, Marcenaro E, et al. Characterization of CD56-/CD16+ natural killer (NK) cells: a highly dysfunctional NK subset expanded in HIV-infected viremic individuals. *Proc Natl Acad Sci USA*. (2005) 102:2886–91. doi: 10.1073/pnas.0409872102
21. Guma M, Cabrera C, Erkizia I, Bofill M, Clotet B, Ruiz L, et al. Human cytomegalovirus infection is associated with increased proportions of NK cells that express the CD94/NKG2C receptor in aviremic HIV-1-positive patients. *J Infect Dis*. (2006) 194:38–41. doi: 10.1086/504719
22. Mela CM, Goodier MR. The contribution of cytomegalovirus to changes in NK cell receptor expression in HIV-1-infected individuals. *J Infect Dis*. (2007) 195:158–9. doi: 10.1086/509811
23. Bjorkstrom NK, Riese P, Heuts F, Andersson S, Fauriat C, Ivarsson MA, et al. Expression patterns of NKG2A, KIR, and CD57 define a process of CD56dim NK-cell differentiation uncoupled from NK-cell education. *Blood*. (2010) 116:3853–64. doi: 10.1182/blood-2010-04-281675
24. Beziat V, Dalgard O, Asselah T, Halfon P, Bedossa P, Boudifa A, et al. CMV drives clonal expansion of NKG2C+ NK cells expressing self-specific KIRs in chronic hepatitis patients. *Eur J Immunol*. (2012) 42:447–57. doi: 10.1002/eji.201141826
25. Djaoud Z, David G, Bressollette C, Willem C, Rettman P, Gagne K, et al. Amplified NKG2C+ NK cells in cytomegalovirus (CMV) infection preferentially express killer cell Ig-like receptor 2DL: functional impact in controlling CMV-infected dendritic cells. *J Immunol*. (2013) 191:2708–16. doi: 10.4049/jimmunol.1301138
26. Manser AR, Scherenschlich N, Thons C, Hengel H, Timm J, Uhrberg M. KIR polymorphism modulates the size of the adaptive NK cell pool in human cytomegalovirus-infected individuals. *J Immunol*. (2019) 203:2301–9. doi: 10.4049/jimmunol.1900423
27. Liu Q, Sun Y, Rihn S, Nolting A, Tsoukas PN, Jost S, et al. Matrix metalloprotease inhibitors restore impaired NK cell-mediated antibody-dependent cellular cytotoxicity in human immunodeficiency virus type 1 infection. *J Virol*. (2009) 83:8705–12. doi: 10.1128/JVI.02666-08
28. Schlums H, Cichocki F, Tesi B, Theorell J, Beziat V, Holmes TD, et al. Cytomegalovirus infection drives adaptive epigenetic diversification of NK cells with altered signaling and effector function. *Immunity*. (2015) 42:443–56. doi: 10.1016/j.immuni.2015.02.008
29. Amiot L, Vu N, Samson M. Immunomodulatory properties of HLA-G in infectious diseases. *J Immunol Res*. (2014) 2014:298569. doi: 10.1155/2014/298569
30. Murdaca G, Contini P, Setti M, Cagnati P, Lantieri F, Indiveri F, et al. Behavior of non-classical soluble HLA class G antigens in human immunodeficiency virus 1-infected patients before and after HAART: comparison with classical soluble HLA-A, -B, -C antigens and potential role in immune-reconstitution. *Clin Immunol*. (2009) 133:238–44. doi: 10.1016/j.clim.2009.08.002
31. Yan WH, Lin A, Chen BG, Chen SY. Induction of both membrane-bound and soluble HLA-G expression in active human cytomegalovirus infection. *J Infect Dis*. (2009) 200:820–6. doi: 10.1086/604733
32. Lee N, Goodlett DR, Ishitani A, Marquardt H, Geraghty DE. HLA-E surface expression depends on binding of TAP-dependent peptides derived from certain HLA class I signal sequences. *J Immunol*. (1998) 160:4951–60.
33. Ulbrecht M, Couturier A, Martinozzi S, Pla M, Srivastava R, Peterson PA, et al. Cell surface expression of HLA-E: interaction with human beta2-microglobulin and allelic differences. *Eur J Immunol*. (1999) 29:537–47. doi: 10.1002/(SICI)1521-4141(199902)29:02<537::AID-IMMU537>3.0.CO;2-6
34. Grimsley C, Ober C. Population genetic studies of HLA-E: evidence for selection. *Human Immunol*. (1997) 52:33–40. doi: 10.1016/S0198-8859(96)00241-8
35. Cohen GB, Gandhi RT, Davis DM, Mandelboim O, Chen BK, Strominger JL, et al. The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity*. (1999) 10:661–71. doi: 10.1016/S1074-7613(00)80065-5
36. Apps R, Del Prete GQ, Chatterjee P, Lara A, Brumme ZL, Brockman MA, et al. HIV-1 Vpu mediates HLA-C downregulation. *Cell Host Microbe*. (2016) 19:686–95. doi: 10.1016/j.chom.2016.04.005
37. Bachtel ND, Umviligihozo G, Pickering S, Mota TM, Liang H, Del Prete GQ, et al. HLA-C downregulation by HIV-1 adapts to host HLA genotype. *PLoS Pathogens*. (2018) 14:e1007257. doi: 10.1371/journal.ppat.1007257
38. Tomasec P, Braud VM, Rickards C, Powell MB, McSharry BP, Gadola S, et al. Surface expression of HLA-E, an inhibitor of natural killer cells,

enhanced by human cytomegalovirus gpUL40. *Science*. (2000) 287:1031. doi: 10.1126/science.287.5455.1031

39. Walters LC, Harlos K, Brackenridge S, Rozbesky D, Barrett JR, Jain V, et al. Pathogen-derived HLA-E bound epitopes reveal broad primary anchor pocket tolerability and conformationally malleable peptide binding. *Nat Commun*. (2018) 9:3137. doi: 10.1038/s41467-018-05459-z
40. Moreno-Cubero E, Ogbe A, Cohen MS, Haynes BF, Borrow P, Peppa D. Subordinate effect of -21M HLA-B dimorphism on NK cell repertoire diversity and function in HIV-1 infected individuals of African origin. *bioRxiv [Preprint]*. (2019) 786392. doi: 10.1101/786392

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fimmu.2020.00314

# High-Throughput *MICA/B* Genotyping of Over Two Million Samples: Workflow and Allele Frequencies

*Anja Klussmeier[1]*, Carolin Massalski[1], Kathrin Putke[1], Gesine Schäfer[1], Jürgen Sauter[2], Daniel Schefzyk[2], Jens Pruschke[2], Jan Hofmann[2], Daniel Fürst[3,4], Raphael Carapito[5], Seiamak Bahram[5], Alexander H. Schmidt[1,2] and Vinzenz Lange[1]*

*[1] DKMS Life Science Lab, Dresden, Germany, [2] DKMS, Tübingen, Germany, [3] Institute of Clinical Transfusion Medicine and Immunogenetics Ulm, German Red Cross Blood Transfusion Service, Baden Wuerttemberg – Hessen, and University Hospital Ulm, Ulm, Germany, [4] Institute of Transfusion Medicine, University of Ulm, Ulm, Germany, [5] Laboratoire d'ImmunoRhumatologie Moléculaire, Plateforme GENOMAX, INSERM UMR_S 1109, LabEx TRANSPLANTEX, Université de Strasbourg, Strasbourg, France*

MICA and MICB are ligands of the NKG2D receptor and thereby influence NK and T cell activity. *MICA/B* gene polymorphisms, expression levels and the amount of soluble MICA/B in the serum have been linked to autoimmune diseases, infections, and cancer. In hematopoietic stem cell transplantation, *MICA* matching between donor and patient has been correlated with reduced acute and chronic graft-vs.-host disease and improved survival. Hence, we developed an extremely cost-efficient high-throughput workflow for genotyping *MICA/B* for newly registered potential stem cell donors. Since mid-2017, we have genotyped over two million samples using NGS amplicon sequencing for *MICA/B* exons 2–5. In donors of German origin, *MICA*008* is the most common *MICA* allele with a frequency of 42.3%. It is followed by *MICA*002* (11.7%) and *MICA*009* (8.8%). The three most common *MICB* alleles are *MICB*005* (43.9%), *MICB*004* (21.7%), and *MICB*002* (18.9%). In general, *MICB* is less diverse than *MICA* and only 6 alleles, instead of 15, account for a cumulative allele frequency of 99.5%. In 0.5% of the samples we observed at least one allele of *MICA* or *MICB* which has so far not been reported to the IPD/IMGT-HLA database. By providing *MICA/B* typed voluntary donors, clinicians become empowered to include *MICA/B* into their donor selection process to further improve unrelated hematopoietic stem cell transplantation.

**Keywords: MICA, MICB, hematopoietic stem cell transplantation, allele, genotyping, next generation sequencing, NGS, high-throughput**

OPEN ACCESS

***Edited by:***
*Martin Maiers,*
*National Marrow Donor Program,*
*United States*

***Reviewed by:***
*Steven Thomas Cox,*
*Anthony Nolan, United Kingdom*
*Marco Andreani,*
*Bambino Gesù Ospedale Pediatrico*
*(IRCCS), Italy*

****Correspondence:***
*Anja Klussmeier*
*klussmeier@dkms-lab.de*

***Specialty section:***
*This article was submitted to*
*Alloimmunity and Transplantation,*
*a section of the journal*
*Frontiers in Immunology*

***Received:*** *20 December 2019*
***Accepted:*** *07 February 2020*
***Published:*** *21 February 2020*

***Citation:***
*Klussmeier A, Massalski C, Putke K, Schäfer G, Sauter J, Schefzyk D, Pruschke J, Hofmann J, Fürst D, Carapito R, Bahram S, Schmidt AH and Lange V (2020) High-Throughput MICA/B Genotyping of Over Two Million Samples: Workflow and Allele Frequencies. Front. Immunol. 11:314. doi: 10.3389/fimmu.2020.00314*

## INTRODUCTION

The *MICA* (MHC class I polypeptide-related sequence A) and *MICB* (MHC class I polypeptide-related sequence B) genes are located between the MHC class I and class III genes inside the human major histocompatibility complex (MHC) (1). Although being highly similar to the classical human leukocyte antigen (*HLA*) genes, they do not present peptides and are not expressed at the surface of human leukocytes but on endothelial cells, fibroblasts, epithelial cells, and tumor cells (2). There they act as ligands for the NKG2D receptor which plays an important role in immune surveillance by activating NK cells and co-stimulating T cell subsets (3, 4). Therefore, the expression of NKG2D ligands is highly regulated and induced by cellular stress (e.g., infection, oxidative stress, transformation).

*MICA* and *MICB* are highly similar and share around 91% of their coding sequence (1). Exon 1 encodes the leader peptide, exons 2, 3, and 4 the three extracellular domains, exon 5 the transmembrane domain and exon 6 the cytoplasmic tail (1, 2, 5). Even though *MICA* and *MICB* do not seem to be as diverse as the conventional *HLA* genes, a large number of distinct alleles have been described: release 3.37.0 of the IPD-IMGT/HLA database contains 109 *MICA* and 47 *MICB* alleles (6). *MICA*008* has been reported to be the most common *MICA* allele with frequencies ranging from 25 to 55% depending on the population. Frequencies above 5% were observed for *MICA*002*, *MICA*009*, *MICA*004*, *MICA*010,* and *MICA*007* in Europeans. In Chinese cohorts, the alleles *MICA*019*, *MICA*027,* and *MICA*045* are also common (7–11). The less diverse *MICB* gene has been predominantly studied in Asian populations. There, the allele *MICB*005* is the most common allele with frequencies of over 50%. It is followed by *MICB*002* and *MICB*004* with frequencies over 10% and *MICB*008* and the null allele *MICB*009N* with frequencies over 5% (10–13).

The most frequent *MICA* allele *MICA*008* differs substantially from most other alleles since it lacks the transmembrane domain due to a frameshift in exon 5. Alleles sharing this feature are also referred to as "A5.1" alleles (14). Their products are bound to the cellular membrane by a GPI-anchor and are frequently released into exosomes thereby triggering a systemic downregulation of the NKG2D receptor on effector cells. Other *MICA* and *MICB* alleles do this to a lesser extent using a soluble form caused by a proteolytic shedding mechanism (15, 16). Since high levels of both forms of soluble MICA and MICB (sMICA/B) have been found in various cancers, the release of MIC proteins is thought to be one cause for cancer immune escape. sMICA/B are therefore considered promising targets for immunotherapy (17–20).

Several studies looked into the general impact of *MICA/B* polymorphisms on different diseases. Especially the *MICA*-129Met/Val dimorphism encoded by the SNP rs1051792 has received attention because it separates the different *MICA* alleles into NKG2D-receptor low (Val)- and high (Met)-affinity binding alleles (21). Health risk associations have been shown for several autoimmune diseases, cancer and viral infections (22–27). Furthermore, matching of *MICA*, including the *MICA*-129 dimorphism, between donor and patient has been correlated with improved outcome of unrelated hematopoietic stem cell transplantation and reduced acute and chronic graft-vs.-host disease (28–32). Because *MICA* is in strong linkage disequilibrium with *HLA-B*, over 90% of 10/10 *HLA*-matched donor/patient pairs are also matched for *MICA* (8, 30). In partially matched cases, in particular in *HLA-B* mismatch situations, *MICA* mismatches are more frequent.

To facilitate further studies on *MICA* and/or *MICB* matching in unrelated hematopoietic stem cell transplantation, we included both genes into our high-throughput genotyping workflow for newly registered potential stem cell donors in 2017. This workflow was initially developed for the six classical *HLA* genes *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQB1,* and *HLA-DPB1* and was then gradually extended to also include *CCR5*, the blood groups *ABO* and *Rh* as well as the several *KIR* genes and *HLA-E* (33–37). Today, this workflow has been applied to genotype over seven million donors, among them more than two million including *MICA* and *MICB.*

# MATERIALS AND METHODS

## Samples

Volunteers from Germany, Poland, UK, USA, Chile and India provided over two million samples to DKMS for their registration as potential stem cell donors between August 2017 and October 2019. We determined *MICA* and *MICB* allele frequencies based on 1,201,896 samples of donors from DKMS Germany who declared to be of German descent. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The described genotyping is within the scope of the consent forms signed at recruitment and performed as genotyping service.

## DNA Isolation and Quantification

The vast majority of samples were provided as buccal swabs (Copan, Brescia, Italy). Few samples were provided as blood. DNA was isolated using the chemagic$^{TM}$ Blood/Swab Kits (PerkinElmer chemagen Technologie GmbH, Baesweiler, Germany) and quantified by fluorescence as described before (36).

## PCR Amplification

*MICA* and *MICB* were amplified in one multiplexed PCR reaction targeting exons 2, 3, and 4/5. The resulting amplicons had lengths between 417 and 480 bp (**Figure 1**). Exons 2 and 3 were amplified as separate amplicons and were completely covered. In contrast, exons 4 and 5 were amplified together as one joined amplicon with primers inside the exons. Therefore, 65 bases at the beginning of exon 4 and 13 bases at the end of exon 5 were not covered. The 8 μl PCR reactions were performed in 384-well plates using FastStart$^{TM}$ Taq DNA Polymerase (Roche, Basel, Switzerland) in its associated buffer system. After amplification, products were pooled with other amplicons of the same sample and subjected to a barcoding/indexing PCR as described previously (33–37).

## Library Preparation and Sequencing

After indexing PCR, 384 barcoded samples were pooled together and purified using SPRIselect beads (BeckmanCoulter, Brea, USA) with a ratio of 0.6:1 beads to DNA and subsequently quantified by qPCR. Equimolar amounts of 10 pools were then combined to a final sequencing library which contained all amplicons from 3,840 donors. The library was denatured and diluted as recommended by Illumina (MiSeq Reagent Kit V2-Reagent Preparation Guide) and loaded at 12.5 pM onto HiSeq flow cells. Paired-end sequencing was performed at 2 × 249 bp using HiSeq Rapid SBS Kits V2 (500 cycles) on HiSeq2500 instruments (Illumina, San Diego, USA) (33–37).

## Genotyping

The neXtype software was extended to support *MICA* and *MICB* genotyping (33, 36). It uses a decision-tree-based algorithm to match the generated *MICA/B* amplicons to known alleles from the official IPD/IMGT-HLA database. Since no known *MICA* amplicon sequence matches a known *MICB* amplicon sequence, reads could be unambiguously assigned to either *MICA* or *MICB*. For more than 95% of the samples neXtype generated correct results with only minor requirements for user interaction. In case of insufficient read coverage, rare or questionable results, a new PCR reaction was initiated from the original DNA. If a low read coverage was limited to exons 4 and 5, trained analysts could decide to generate a result based on exons 2 and 3 only. Genotyping results were finally exported using the GL string format (38).

## Frequency Analysis of *MICA* and *MICB* Alleles

*MICA* and *MICB* genotyping results of 1,201,896 samples of German origin were analyzed based on the first field, which identifies the unique MICA and MICB proteins. Homozygous genotyping results were counted as two alleles. Allele groups which could not be distinguished due to missing sequencing information were reported by a representative allele which was marked with a hash symbol (#) (**Table 1**). For samples with phasing ambiguities, the probability of each possible result was calculated based on the allele frequencies of unambiguously typed samples. According to these probabilities, counts were added to the different alleles. To verify rare allele calls, all alleles observed <50 times were reconfirmed in at least two samples.

# RESULTS

## High-Throughput *MICA/B* Genotyping

### Assay Validation and Performance

For assay validation, we exchanged DNA from 95 samples with two labs with established workflows for *MICA* or *MICB* genotyping (*MICA*: Institute of Clinical Transfusion Medicine and Immunogenetics Ulm, Germany; *MICB*: Laboratoire d'ImmunoRhumatologie Moléculaire, Strasbourg, France). For *MICA*, we additionally used the UCLA *MICA* Panel Set (UCLA Immunogenetic Center, USA), which consists of 24 samples with diverse combinations of *MICA* alleles. The results obtained from our newly established workflow were 100% concordant with the reference genotypes for both *MICA* and *MICB* (**Supplementary File 1**). Subsequently, *MICA/B* genotyping was included into our standard genotyping workflow in August 2017 and applied for all newly registered donors. So far, we have generated *MICA/B* genotyping data for over two million samples, on average more than 20,000 samples per week. Because *MICA/B* amplicons are pooled with the *HLA* amplicons directly after the initial PCR, additional costs for genotyping *MICA/B* are minor and reflect the costs for one 8 μl PCR reaction, sequencing and data analysis. We are targeting an average coverage of 1,000 reads per locus and exon corresponding to a total of 6,000 reads for *MICA/B* with associated costs of about 10 cents per sample for sequencing. This efficient strategy makes it feasible to genotype every newly registered donor for *MICA/B*.

### Resolution and Ambiguities

Our *MICA/B* genotyping workflow targets and amplifies exons 2 and 3 separately and most of exons 4 and 5 using a combined amplicon (**Figure 1**). Consequently, exons 1 and 6 and 78 bases of exons 4 and 5 are not sequenced. This amplification strategy promised a good genotyping resolution while being highly cost-efficient. *MICA/B* exons 2, 3, and 5 were considered mandatory because they encode the receptor-interacting domains or define *MICA*008*-like alleles. Expansion of the exon 5 amplicon made it possible to also include most of exon 4. Exons 1 and 6 encode a leader peptide and the cytoplasmic tail. As these regions do not encode extracellular domains of the proteins and are characterized by a lower diversity they were not included in the genotyping strategy. However, some alleles may only be differentiated by sequence features within one of the not covered regions. For example, SNPs in exon 6 are the only way to distinguish *MICA*010* from *MICA*069* or *MICA*009:01* from *MICA*049*. *MICA*009:02*, on the other hand, can be unambiguously genotyped because it differs from *MICA*049* and

**TABLE 1** | Overview of ambiguous genotyping results.

| Allele group | Alleles |
|---|---|
| ***MICA*** | |
| *MICA*009#* | *MICA*009*, *MICA*049* |
| *MICA*010#* | *MICA*010*, *MICA*065*, *MICA*069* |
| *MICA*027#* | *MICA*027*, *MICA*048* |
| ***MICB*** | |
| *MICB*004#* | *MICB*004*, *MICB*028* |
| *MICB*005#* | *MICB*003*, *MICB*005*, *MICB*006*, *MICB*010* |
| *MICB*014#* | *MICB*014*, *MICB*015* |

*Alleles which cannot be distinguished from each other by the workflow are combined in an allele group marked with a hash symbol (#).*

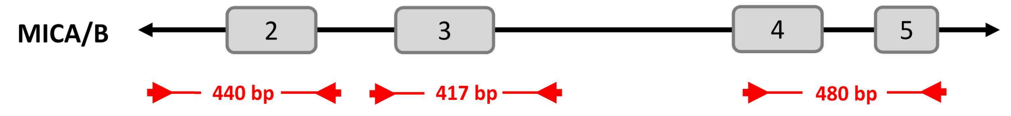


**FIGURE 1** | Primer locations and PCR amplification products for exons 2–5 of *MICA/B*. Primers (arrows) bind to both *MICA* and *MICB* and generate three amplicons per gene in one PCR reaction. Product lengths are between 417 and 480 bp. Note that not all bases of exons 4 and 5 are covered.

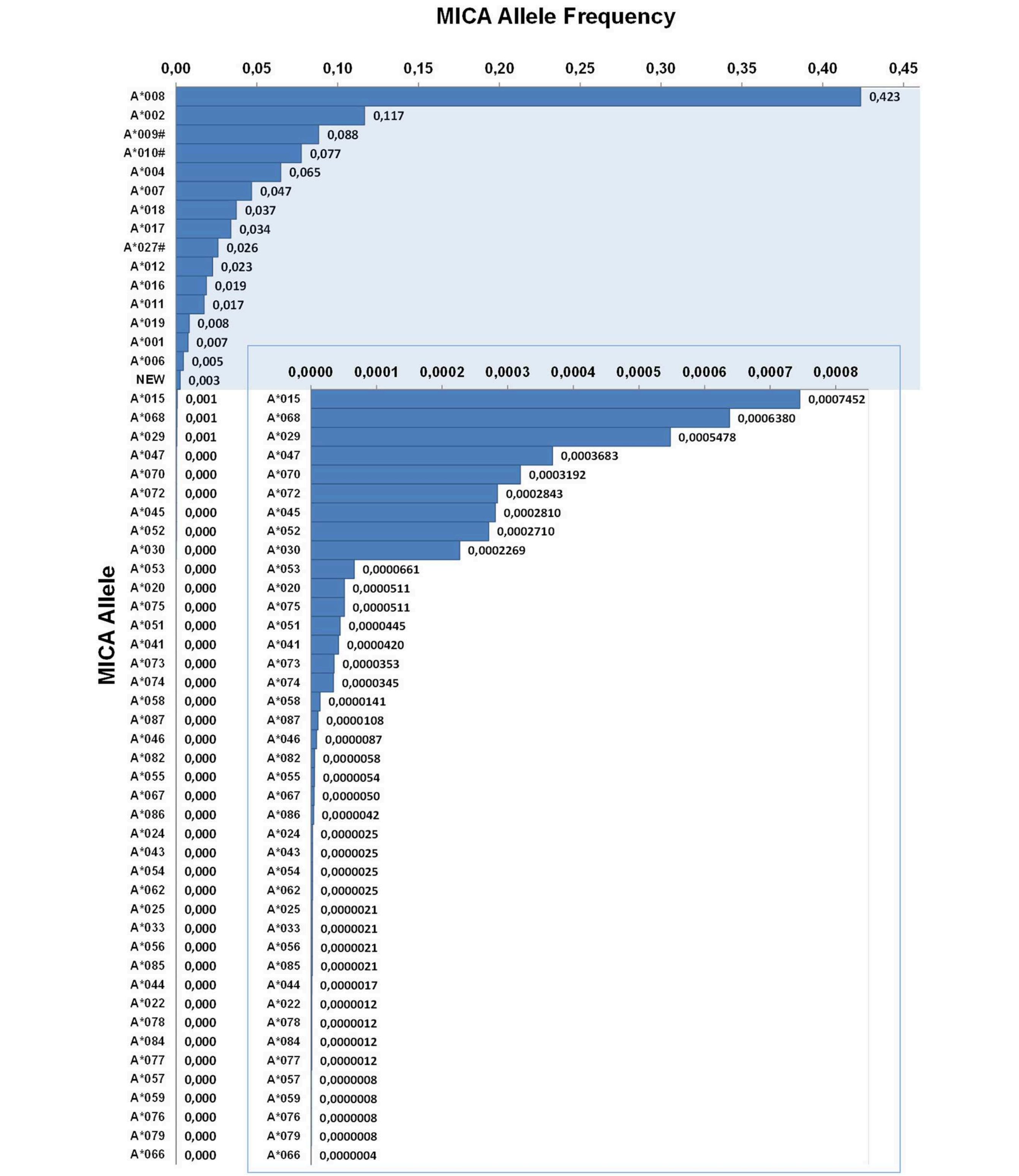


**FIGURE 2 |** Allele frequencies of *MICA*. First-field-resolution allele frequencies are based on 1,201,896 samples from donors of German descent. Alleles contributing to a cumulative allele frequency of 99.5% are shown against a colored background and allele frequencies below 0.003 are additionally plotted in an inlay. If ambiguities exist, allele groups are used (#) and the ambiguity is described in **Table 1**.

its synonymous allele *MICA*009:01* in exon 3 (**Table 1**) (14). Due to the primer location inside exon 4 our workflow also cannot distinguish between *MICA*10* and *MICA*065*.

For *MICB*, the most common allele *MICB*005:02* cannot be distinguished from *MICB*003*, *MICB*006,* and *MICB*010*, while other variants of *MICB*005* can be distinguished. Likewise, the pairs *MICB*004* and *MICA*028* or *MICB*014* and *MICA*015* cannot be resolved (**Table 1**).

In addition to the ambiguities caused by missing sequence information, we encounter phasing ambiguities. They occur because the sequences of short amplicons cannot be phased if the targeted regions are not overlapping. As a consequence, some observed sequence combinations can be explained by more than one allele pair. In our workflow, phasing ambiguities occur in 3% of *MICA* and 24% of *MICB* samples. In over 99.9% of those cases, however, one possibility can statistically be ruled out since the combination of two rare alleles would be highly unlikely if the other option includes two common alleles. This is in contrast to *HLA* genotyping where some important phasing ambiguities cannot be solved statistically. For example, the most common *MICB* phasing ambiguity result is either the combination *MICB*002* and *MICB*005#* or the combination *MICB*018* and *MICB*019* [GL-String notation: *MICB*002+MICB*005#|MICB*018+MICB*019* (38)]. Based on the allele frequencies determined in this study, the likelihood of the allele combination *MICB*002+MICB*005#* is 0.039. In contrast, the likelihood of *MICB*018+MICB*019* is only $1.9 \times 10^{-8}$. Hence, *MICB*018+MICB*019* would be expected to occur only once in 2.08 million samples with the given phasing result. In our dataset of 1,201,896 samples, 182,383 samples have the result *MICB*002+MICB*005#|MICB*018+MICB*019*. Now, by claiming that *MICB*002+MICB*005#* is always the correct result, we are making only one wrong call in 13.7 million genotyped samples. Therefore, we have disregarded the highly unlikely combinations of rare alleles in our allele frequency calculations. This is not expected to introduce a relevant error. In contrast, disregarding all samples with phasing results altogether would substantially distort the results since the phasing events predominantly involve certain alleles.

### Novel Alleles

We encounter novel *MICA* or *MICB* alleles in 0.5% of the samples, resulting in the observation of ∼100 novel alleles per week (recurrences included). They are automatically flagged by the genotyping software and trigger a new PCR reaction from the original sample for verification. In general, the novel alleles fall into two categories: Novel sequences or novel combinations of previously reported exonic sequences. The task to characterize them in full length and submit the sequences to IPD/IMGT-HLA is currently in progress.

## *MICA* Allele Frequencies

*MICA* allele frequencies were calculated on 1,201,896 samples of German descent (**Figure 2**). These samples represent more than 50% of our genotyped samples and were therefore the largest ethnically defined population available. With a frequency of 42.3%, the allele *MICA*008* is the most frequent *MICA* allele in Germany. It is followed by the alleles *MICA*002* (11.7%), *MICA*009#* (8.8%), *MICA*010#* (7.7%), and *MICA*004* (6.5%). The 15 most common alleles account for a cumulative allele frequency of 99.5%. The other 41 alleles observed in the German dataset account for the remaining 0.5%. We further identified six *MICA* alleles (*MICA*035*, *MICA*037*, *MICA*038*, *MICA*040*, *MICA*060,* and *MICA*064N*) with very low frequencies in samples not of German origin. Despite the huge sample size, we have never observed the remaining 18 alleles contained in the IPD-IMGT/HLA database (release 3.37.0) (**Table 2**).

**TABLE 2** | *MICA/B* alleles described in IPD/IMGT-HLA release 3.37.0, but never observed in our cohort of over two million samples.

| | |
|---|---|
| *MICA* | *MICA*005*, *MICA*013*, *MICA*014*, *MICA*023*, *MICA*026*, *MICA*028*, *MICA*031*, *MICA*032*, *MICA*034*, *MICA*036*, *MICA*039*, *MICA*042*, *MICA*050*, *MICA*061*, *MICA*063N*, *MICA*065*, *MICA*081*, *MICA*083* |
| *MICB* | *MICB*001*, *MICB*011*, *MICB*016*, *MICB*022*, *MICB*030*, *MICB*032* |

## *MICB* Allele Frequencies

*MICB* allele frequencies were calculated based on the same sample cohort used for *MICA* (**Figure 3**). With a frequency of 43.9%, *MICB*005#* is by far the most frequent allele in Germany. However, since our workflow cannot distinguish all *MICB*005* variants from *MICB*003*, *MICB*006,* and *MICB*010*, the true frequency of *MICB*005* might be lower (**Table 1**). In our samples, *MICB*005#* is followed by *MICB*004#*, *MICB*002,* and *MICB*008*, having frequencies of 21.7 18.9, and 11.0%, respectively. Together with *MICB*014#* (2.2%) and *MICB*013* (1.4%) they account for a cumulative allele frequency of 99.5%. 14 other alleles have been detected in the German cohort. *MICB*007* has only been identified in a few samples of non-German origin. We have never observed the six remaining alleles described in the IPD-IMGT/HLA database (release 3.37.0) (**Table 2**).

# DISCUSSION

The regulation of NK/T cell activation is an elaborate interplay between several receptors and their associated ligands. To further add another layer of complexity, receptors like KIR or ligands like MICA/B exist in a variety of distinct alleles with varying effects on NK/T cell activity (6, 39). A comprehensive sequencing study of the MHC complex indicated that the sequence of *MICA* is more diverse than the sequence of *HLA-DQB1* or *HLA-DPB1*, but the number of named *MICA* alleles is much lower (6, 10). And even though MICA/B do not present antigenic peptides like the classical *HLA* class I genes, matching of *MICA/B* between patient and donor has been reported to improve outcome and reduce acute and chronic graft-vs.-host disease in hematopoietic stem cell transplantation, especially in partially matched scenarios (30, 31, 40). Translation of these findings into clinical practice is, amongst others, hampered by the lack of *MICA/B* genotyping data. Hence, we present a workflow to genotype both *MICA* and *MICB* with a mean throughput of over 20,000 samples

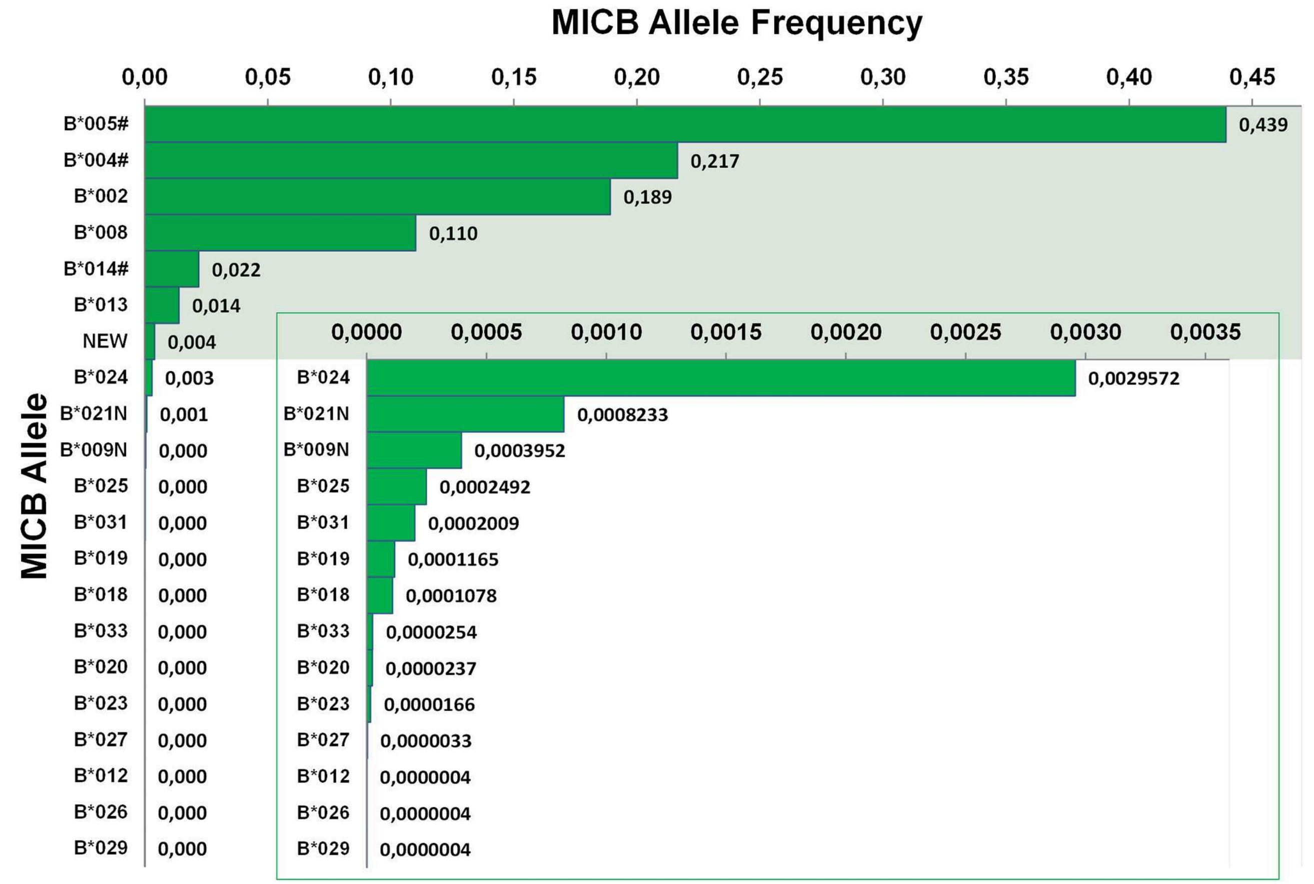


**FIGURE 3 |** Allele frequencies of *MICB*. First-field-resolution allele frequencies are based on 1,201,896 samples from donors of German descent. Alleles contributing to a cumulative allele frequency of 99.5% are shown against a colored background and allele frequencies below 0.004 are additionally plotted in an inlay. If ambiguities exist, allele groups are used (#) and the ambiguity is described in **Table 1**.

per week. To date, we have processed more than two million donor samples.

Based on 1.2 million samples of German origin we identified *MICA**008* as the most common *MICA* allele (42.3%), followed by *MICA**002* (11.7%) and *MICA**009#* (8.8%). This is concordant to previous studies which present allele frequencies between 43 and 55% for *MICA**008*, 8–14% for *MICA**002* and 4–8% for *MICA**009* in European/American populations (7–9). Although *MICA**008* is also the most common allele in China, with a frequency of about 25% it is far less abundant than in European/American populations (10, 11, 41). Since *MICA**008* and other rare alleles bearing the A5.1 microsatellite marker are more prone to produce sMICA than other alleles, they are more effective in inactivating NKG2D and NK/T cell activity (15). Therefore, these alleles might contribute to the disease prevalence in different populations. Indeed, A5.1-carriers have been associated with an increased risk for several types of cancer and higher levels of sMICA seem to have a negative prognostic value for tumor patient survival (18, 27, 42–44). To reactivate a patient's NK cells, the reduction of soluble NKG2D ligands is a promising approach. Current strategies comprise the inhibition of enzymes responsible for shedding as well as blocking the cleavage sites with therapeutic antibodies. Most likely, the efficacy of some of these new drugs will be limited to certain *MICA/B* alleles which increases the need for reliable genotyping (18, 45).

*MICB* is less diverse than *MICA*. The most common allele *MICB**005#* was detected at 43.9% allele frequency in the German population. However, given the incomplete sequence coverage, our workflow cannot distinguish *MICB**003*, *MICB**005*, *MICB**006*, and *MICB**010*. Studies on Asian cohorts report allele frequencies of at least 55% for *MICB**005*, 3% for *MICB**003* and no observations of *MICB**006* or *MICB**010* (10, 11, 13). Limited full gene analysis of 51 samples with *MICB**005#* pre-typing results indicated a similar distribution in our dataset (data not shown).

The *MICB**003/005:02* ambiguity with its distinguishing bases at the beginning of exon 4 and in exon 6 is one case in which our workflow cannot differentiate between two presumably common alleles. However, an amplicon of at least 530 bp would be necessary to include the SNP at the beginning of exon 4 and to not lose sequencing information for the microsatellite region in *MICA* exon 5. Since this exceeds Illumina's 2 × 250 bp read length, bases at the end of exon 4 would not be sequenced, thereby creating other ambiguities. Consequently, to clearly distinguish between *MICB*005:02* and *MICB*003* a separate fourth PCR amplicon would be required. But given the lack of clinical data for the relevance of regions outside exons 2, 3, and 5, one might wonder if a higher resolution for *MICA/B* genotyping is necessary. In *HLA* genotyping transplantation compatible allele groups have been defined (G or P Codes) combining all alleles harboring the same sequence across the antigen recognition domain (2, 46, 47). For *MICA/B*, there is no similar system yet. Consequently, we do not think that it is proportionate to increase the sequencing costs for all samples without further evidence of the clinical importance of remaining ambiguities. For individual samples, genotyping results with three-field resolution can be generated using long-read sequencing technologies (48). Moreover, our amplicon strategy does not include the 5′ and 3′ UTRs of *MICA/B* which contain additional polymorphic positions (49, 50). Some of them influence (s)MICA/B expression which varies between different alleles (18, 51–53). However, to the best of our knowledge, there are no studies, which address the effects of donor *MICA/B* variations outside the exons in hematopoietic stem cell transplantation.

Although we genotyped over two million samples, we have not encountered some of the *MICA/B* alleles described in the IPD/IMGT-HLA database (**Table 2**). This may be due to several reasons. First of all, the majority of our samples are of European origin. Therefore, we might lack rare alleles occurring predominantly in other ethnicities. One example is *MICB*032* which was originally isolated from an Uyghur individual (54). In other cases, initial submissions to IPD/IMGT-HLA could be erroneous. This might especially be true for the alleles that have never been independently confirmed. For example, all heterozygous positions defining *MICA*005* or *MICA*013* also occur in one of the two most common alleles *MICA*008* and *MICA*002*. If those positions were not correctly phased during Sanger sequence analysis, *MICA*005* and *MICA*013* could have been erroneously reported. However, the sequencing of cloned PCR fragments should have prevented such errors (1, 55). Other not observed alleles, like *MICA*081*, *MICB*011*, *MICB*016,* or *MICB*022*, differ from more common alleles in only one position (56, 57). While this may reflect sequencing errors, it is more likely that the more recent submissions represent very low frequency observations as we discover on a daily basis. However, for the individual allele this may only be resolved by resequencing the original DNA which is often no longer available.

In conclusion, our workflow demonstrates that upfront *MICA/B* genotyping for potential stem cell donors can be performed with only minor increases in expenses and workload. So far, *MICA/B* informed donor selection has not yet found widespread application in clinical practice. Clearly, additional confirmatory studies would be worthwhile. However, the availability of genotyping information remains a major hurdle for the translation of new markers into clinical practice. With the *MICA/B* genotyping of millions of donors we provide that data to facilitate *MICA/B* informed donor selection.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

KP developed and tested the primer set. JS, DS, JP, and JH developed and implemented the genotyping algorithm. DF, RC, and SB performed genotyping for reference samples. AK, CM, GS, and JS analyzed frequency data. AK prepared figures and tables. AK and VL wrote the manuscript. AK, AS, and VL conceived and supervised the work. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

RC and SB were supported by the Agence Nationale de la Recherche (ANR)—ANR-11-LABX-0070_TRANSPLANTEX and MSD-Avenir grant AUTOGEN.

## ACKNOWLEDGMENTS

We are grateful to all members of the DKMS Life Science Lab for their dedicated daily work that was fundamental for the analysis of all the donor samples.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00314/full#supplementary-material

## REFERENCES

1. Bahram S, Bresnahan M, Geraghty DE, Spies T. A second lineage of mammalian major histocompatibility complex class I genes. *Proc Natl Acad Sci USA*. (1994) 91:6259–63. doi: 10.1073/pnas.91.14.6259
2. Risti M, Bicalho MD. MICA and NKG2D: is there an impact on kidney transplant outcome? *Front Immunol*. (2017) 8:179. doi: 10.3389/fimmu.2017.00179
3. Bauer S, Groh V, Wu J, Steinle A, Phillips JH, Lanier LL, et al. Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. *Science*. (1999) 285:727–9. doi: 10.1126/science.285.5428.727
4. Glienke J, Sobanov Y, Brostjan C, Steffens C, Nguyen C, Lehrach H, et al. The genomic organization of NKG2C, E, F, and D receptor genes in the human natural killer gene complex. *Immunogenetics*. (1998) 48:163–73. doi: 10.1007/s002510050420

5. Li P, Morris DL, Willcox BE, Steinle A, Spies T, Strong RK. Complex structure of the activating immunoreceptor NKG2D and its MHC class I–like ligand MICA. *Nat Immunol*. (2001) 2:443–51. doi: 10.1038/87757
6. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucl Acids Res*. (2015) 43:D423-31.
7. Petersdorf EW, Shuler KB, Longton GM, Spies T, Hansen JA. Population study of allelic diversity in the human MHC class I-related MIC-A gene. *Immunogenetics*. (1999) 49:605–12. doi: 10.1007/s002510050655
8. Gao X, Single RM, Karacki P, Marti D, O'Brien SJ, Carrington M. Diversity of MICA and linkage disequilibrium with HLA-B in two North American populations. *Hum Immunol*. (2006) 67:152–8. doi: 10.1016/j.humimm.2006.02.009
9. Ahmad T, Marshall SE, Mulcahy-Hawes K, Orchard T, Crawshaw J, Armuzzi A, et al. High resolution MIC genotyping: design and application to the investigation of inflammatory bowel disease susceptibility. *Tissue Antigens*. (2002) 60:164–79. doi: 10.1034/j.1399-0039.2002.600207.x
10. Zhou F, Cao H, Zuo X, Zhang T, Zhang X, Liu X, et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat Genet*. (2016) 48:740–6. doi: 10.1038/ng.3576
11. Wang W, Tian W, Zhu F, Li L, Cai J, Wang F, et al. MICA Gene Deletion in 3411 DNA Samples from Five Distinct Populations in Mainland China and Lack of Association with Nasopharyngeal Carcinoma (NPC) in a Southern Chinese Han population. *Ann Hum Genet*. (2016) 80:319–26. doi: 10.1111/ahg.12175
12. Ying Y, He Y, Tao S, Han Z, Wang W, Chen N, et al. Distribution of MICB diversity in the Zhejiang Han population: PCR sequence-based typing for exons 2–6 and identification of five novel MICB alleles. *Immunogenetics*. (2013) 65:485–92. doi: 10.1007/s00251-013-0699-4
13. Cha CH, Sohn YH, Oh HB, Ko SY, Cho MC, Kwon OJ. MICB polymorphisms and haplotypes with MICA and HLA alleles in Koreans. *Tissue Antigens*. (2011) 78:38–44. doi: 10.1111/j.1399-0039.2011.01694.x
14. Frigoul A, Lefranc M-P. MICA: standardized IMGT allele nomenclature, polymorphisms and diseases. In: Pandalai SG, editor. *Recent Research Developments in Human Genetics*, Vol. 3. Trivandrum: Research Signpost (2005). p. 95–145.
15. Ashiru O, Boutet P, Fernández-Messina L, Agüera-González S, Skepper JN, Valés-Gómez M, et al. Natural killer cell cytotoxicity is suppressed by exposure to the human NKG2D ligand MICA*008 that is shed by tumor cells in exosomes. *Cancer Res*. (2010) 70:481–9. doi: 10.1158/0008-5472.CAN-09-1688
16. Ashiru O, López-Cobo S, Fernández-Messina L, Pontes-Quero S, Pandolfi R, Reyburn HT, et al. A GPI anchor explains the unique biological features of the common NKG2D-ligand allele MICA*008. *Biochem J*. (2013) 454:295–302. doi: 10.1042/BJ20130194
17. Nückel H, Switala M, Sellmann L, Horn PA, Dürig J, Dührsen U, et al. The prognostic significance of soluble NKG2D ligands in B-cell chronic lymphocytic leukemia. *Leukemia*. (2010) 24:1152–9. doi: 10.1038/leu.2010.74
18. Schmiedel D, Mandelboim O. NKG2D ligands-critical targets for cancer immune escape and therapy. *Front Immunol*. (2018) 9:2040. doi: 10.3389/fimmu.2018.02040
19. Duan S, Guo W, Xu Z, He Y, Liang C, Mo Y, et al. Natural killer group 2D receptor and its ligands in cancer immune escape. *Mol Cancer*. (2019) 18:29. doi: 10.1186/s12943-019-0956-8
20. de Andrade LF, Tay RE, Pan D, Luoma AM, Ito Y, Badrinath S, et al. Antibody-mediated inhibition of MICA and MICB shedding promotes NK cell–driven tumor immunity. *Science*. (2018) 359:1537–42. doi: 10.1126/science.aao0505
21. Steinle A, Li P, Morris DL, Groh V, Lanier LL, Strong RK, et al. Interactions of human NKG2D with its ligands MICA, MICB, and homologs of the mouse RAE-1 protein family. *Immunogenetics*. (2001) 53:279–87. doi: 10.1007/s002510100325
22. Zuo J, Mohammed F, Moss P. The Biological influence and clinical relevance of polymorphism within the NKG2D ligands. *Front Immunol*. (2018) 9:1820. doi: 10.3389/fimmu.2018.01820
23. Pollock RA, Chandran V, Pellett FJ, Thavaneswaran A, Eder L, Barrett J, et al. The functional MICA-129 polymorphism is associated with skin but not joint manifestations of psoriatic disease independently of HLA-B and HLA-C. *Tissue Antigens*. (2013) 82:43–7. doi: 10.1111/tan.12126
24. Tong HV, Toan NL, Song LH, Bock CT, Kremsner PG, Velavan TP. Hepatitis B virus-induced hepatocellular carcinoma: functional roles of MICA variants. *J Viral Hepat*. (2013) 20:687–98. doi: 10.1111/jvh.12089
25. Isernhagen A, Malzahn D, Bickeböller H, Dressel R. Impact of the MICA-129Met/Val dimorphism on NKG2D-mediated biological functions and disease risks. *Front Immunol*. (2016) 7:588. doi: 10.3389/fimmu.2016.00588
26. Chen E, Chen C, Chen F, Yu P, Lin L. Positive association between MIC gene polymorphism and tuberculosis in Chinese population. *Immunol Lett*. (2019) 213:62–9. doi: 10.1016/j.imlet.2019.07.008
27. Carapito R, Gottenberg JE, Kotova I, Untrau M, Michel S, Naegely L, et al. A new MHC-linked susceptibility locus for primary Sjögren's syndrome: MICA. *Hum Mol Genet*. (2017) 26:2565–76. doi: 10.1093/hmg/ddx135
28. Isernhagen A, Malzahn D, Viktorova E, Elsner L, Monecke S, von Bonin F, et al. The MICA-129 dimorphism affects NKG2D signaling and outcome of hematopoietic stem cell transplantation. *EMBO Mol Med*. (2015) 7:1480–502. doi: 10.15252/emmm.201505246
29. Parmar S, Del Lima M, Zou Y, Patah PA, Liu P, Cano P, et al. Donor-recipient mismatches in MHC class I chain-related gene A in unrelated donor transplantation lead to increased incidence of acute graft-versus-host disease. *Blood*. (2009) 114:2884–7. doi: 10.1182/blood-2009-05-223172
30. Fuerst D, Neuchel C, Niederwieser D, Bunjes D, Gramatzki M, Wagner E, et al. Matching for the MICA-129 polymorphism is beneficial in unrelated hematopoietic stem cell transplantation. *Blood*. (2016) 128:3169–76. doi: 10.1182/blood-2016-05-716357
31. Carapito R, Jung N, Kwemou M, Untrau M, Michel S, Pichot A, et al. Matching for the nonconventional MHC-I MICA gene significantly reduces the incidence of acute and chronic GVHD. *Blood*. (2016) 128:1979–86. doi: 10.1182/blood-2016-05-719070
32. Carapito R, Aouadi I, Ilias W, Bahram S. Natural Killer Group 2, Member D/NKG2D ligands in hematopoietic cell transplantation. *Front Immunol*. (2017) 8:368. doi: 10.3389/fimmu.2017.00368
33. Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*. (2014) 15:63. doi: 10.1186/1471-2164-15-63
34. Schöfl G, Lang K, Quenzel P, Böhme I, Sauter J, Hofmann JA, et al. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC Genomics*. (2017) 18:161. doi: 10.1186/s12864-017-3575-z
35. Lang K, Wagner I, Schöne B, Schöfl G, Birkner K, Hofmann JA, et al. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genomics*. (2016) 17:374. doi: 10.1186/s12864-016-2687-1
36. Wagner I, Schefzyk D, Pruschke J, Schöfl G, Schöne B, Gruber N, et al. Allele-Level KIR genotyping of more than a million samples: workflow, algorithm, and observations. *Front Immunol*. (2018) 9:2843. doi: 10.3389/fimmu.2018.02843
37. Solloch UV, Lang K, Lange V, Böhme I, Schmidt AH, Sauter J. Frequencies of gene variant CCR5-Δ32 in 87 countries based on next-generation sequencing of 1.3 million individuals sampled from 3 national DKMS donor centers. *Hum Immunol*. (2017) 78:710–7. doi: 10.1016/j.humimm.2017.10.001
38. Milius RP, Mack SJ, Hollenbach JA, Pollack J, Heuer ML, Gragert L, et al. Genotype List String: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens*. (2013) 82:106–12. doi: 10.1111/tan.12150
39. Marsh SGE, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Hum Immunol*. (2003) 64:648–54. doi: 10.1016/S0198-8859(03)00067-3
40. Carapito R, Jung N, Untrau M, Michel S, Pichot A, Giacometti G, et al. Matching of MHC Class I chain-related genes a and B is associated with reduced incidence of severe acute Graft-Versus-Host disease after unrelated hematopoietic stem cell transplantation. *Blood*. (2014) 124:664. doi: 10.1182/blood.V124.21.664.664
41. Tian W, Cai JH, Wang F, Li LX. MICA polymorphism in a northern Chinese Han population: the identification of a new MICA allele, MICA*059. *Hum Immunol*. (2010) 71:423–7. doi: 10.1016/j.humimm.2010.01.025
42. Chen D, Juko-Pecirep I, Hammer J, Ivansson E, Enroth S, Gustavsson I, et al. Genome-wide association study of susceptibility loci for cervical cancer. *J Natl Cancer Inst*. (2013) 105:624–33. doi: 10.1093/jnci/djt051

43. Jiang X, Zou Y, Huo Z, Yu P. Association of major histocompatibility complex class I chain-related gene A microsatellite polymorphism and hepatocellular carcinoma in South China Han population. *Tissue Antigens*. (2011) 78:143–7. doi: 10.1111/j.1399-0039.2011.01693.x
44. Onyeaghala G, Lane J, Pankratz N, Nelson HH, Thyagarajan B, Walcheck B, et al. Association between MICA polymorphisms, s-MICA levels, and pancreatic cancer risk in a population-based case-control study. *PLoS ONE*. (2019) 14:e0217868. doi: 10.1371/journal.pone.0217868
45. Lombana TN, Matsumoto ML, Berkley AM, Toy E, Cook R, Gan Y, et al. High-resolution glycosylation site-engineering method identifies MICA epitope critical for shedding inhibition activity of anti-MICA antibodies. *MAbs*. (2019) 11:75–93. doi: 10.1080/19420862.2018.1532767
46. HLA Nomenclature @ hla.alleles.org [Internet]. Available online at: http://hla.alleles.org/alleles/g_groups.html (accessed November 29, 2019).
47. HLA Nomenclature @ hla.alleles.org [Internet]. Available online at: http://hla.alleles.org/alleles/p_groups.html (accessed November 29, 2019).
48. Albrecht V, Zweiniger C, Surendranath V, Lang K, Schöfl G, Dahl A, et al. Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles. *HLA*. (2017) 90:79–87. doi: 10.1111/tan.13057
49. Cox ST, Madrigal JA, Saudemont A. Diversity and characterization of polymorphic 5′ promoter haplotypes of MICA and MICB genes. *Tissue Antigens*. (2014) 84:293–303. doi: 10.1111/tan.12400
50. Cox ST, Hernandez D, Danby R, Turner TR, Madrigal JA. Diversity and characterisation of polymorphic 3’ untranslated region haplotypes of MICA and MICB genes. *HLA*. (2018) 92:392–402. doi: 10.1111/tan.13434
51. Rodríguez-Rodero S, González S, Rodrigo L, Fernández-Morera JL, Martínez-Borra J, López-Vázquez A, et al. Transcriptional regulation of MICA and MICB: a novel polymorphism in MICB promoter alters transcriptional regulation by Sp1. *Eur J Immunol*. (2007) 37:1938–53. doi: 10.1002/eji.200737031
52. Lo PHY, Urabe Y, Kumar V, Tanikawa C, Koike K, Kato N, et al. Identification of a functional variant in the MICA promoter which regulates MICA expression and increases HCV-related hepatocellular carcinoma risk. *PLoS ONE*. (2013) 8:e61279. doi: 10.1371/journal.pone.0061279
53. Shi C, Li H, Couturier JP, Yang K, Guo X, He D, et al. Allele specific expression of MICA variants in human fibroblasts suggests a pathogenic mechanism. *Open Rheumatol J*. (2015) 9:60–4. doi: 10.2174/1874312901409010060
54. Alleles Report < IMGT/HLA < IPD < EMBL-EBI [Internet]. [cited 2019 Dec 17]. Available online at: https://www.ebi.ac.uk/cgi-bin/ipd/imgt/hla/get_allele.cgi?MICB^\ast032 (accessed November 29, 2019).
55. Fodil N, Laloux L, Wanner V, Pellet P, Hauptmann G, Mizuki N, et al. Allelic repertoire of the humanMHC class IMICA gene. *Immunogenetics*. (1996) 44:351–7. doi: 10.1007/BF02602779
56. Schroeder M, Elsner HA, Kim TD, Blasczyk R. Eight novel MICB alleles, including a null allele, identified in gastric MALT lymphoma patients. *Tissue Antigens*. (2004) 64:276–80. doi: 10.1111/j.1399-0039.2004.00286.x
57. Visser CJT, Tilanus MGJ, Schaeffer V, Tatari Z, Tamouza R, Janin A, et al. Sequencing-based typing reveals six novel MHC class I chain-related gene B (MICB) alleles. *Tissue Antigens*. (1998) 51:649–52. doi: 10.1111/j.1399-0039.1998.tb03008.x

**Conflict of Interest:** AK, CM, GS, KP, AS, and VL are members of the DKMS Life Science Lab which offers commercial genotyping services.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

ORIGINAL RESEARCH
published: 12 March 2020
doi: 10.3389/fimmu.2020.00429

# Estimation of German KIR Allele Group Haplotype Frequencies

*Ute V. Solloch*[1]*, *Daniel Schefzyk*[1], *Gesine Schäfer*[2], *Carolin Massalski*[2], *Maja Kohler*[2], *Jens Pruschke*[1], *Annett Heidl*[2], *Johannes Schetelig*[1,3], *Alexander H. Schmidt*[1,2], *Vinzenz Lange*[2] *and Jürgen Sauter*[1]

[1] DKMS, Tübingen, Germany, [2] DKMS Life Science Lab, Dresden, Germany, [3] University Hospital Carl Gustav Carus, Dresden, Germany

OPEN ACCESS

**Edited by:**
Martin Maiers,
National Marrow Donor Program,
United States

**Reviewed by:**
Lisbeth Guethlein,
Stanford University, United States
Carlos Vilches,
Hospital Universitario Puerta de Hierro
Majadahonda, Spain
Lihua Hou,
Georgetown University, United States

***Correspondence:**
Ute V. Solloch
solloch@dkms.de

**Specialty section:**
This article was submitted to
Alloimmunity and Transplantation,
a section of the journal
Frontiers in Immunology

**Received:** 14 November 2019
**Accepted:** 25 February 2020
**Published:** 12 March 2020

**Citation:**
Solloch UV, Schefzyk D, Schäfer G,
Massalski C, Kohler M, Pruschke J,
Heidl A, Schetelig J, Schmidt AH,
Lange V and Sauter J (2020)
Estimation of German KIR Allele
Group Haplotype Frequencies.
Front. Immunol. 11:429.
doi: 10.3389/fimmu.2020.00429

The impact of the highly polymorphic Killer-cell immunoglobulin-like receptor (*KIR*) gene cluster on the outcome of hematopoietic stem cell transplantation (HCST) is subject of current research. To further understand the involvement of this gene family into Natural Killer (NK) cell-mediated graft-versus-leukemia reactions, knowledge of haplotype structures, and allelic linkage is of importance. In this analysis, we estimate population-specific *KIR* haplotype frequencies at allele group resolution in a cohort of $n = 458$ German families. We addressed the polymorphism of the *KIR* gene complex and phasing ambiguities by a combined approach. Haplotype inference within first-degree family relations allowed us to limit the number of possible diplotypes. Structural restriction to a pattern set of 92 previously described *KIR* copy number haplotypes further reduced ambiguities. *KIR* haplotype frequency estimation was finally accomplished by means of an expectation-maximization algorithm. Applying a resolution threshold of ½ *n*, we were able to identify a set of 551 *KIR* allele group haplotypes, representing 21 *KIR* copy number haplotypes. The haplotype frequencies allow studying linkage disequilibrium in two-locus as well as in multi-locus analyses. Our study reveals associations between *KIR* haplotype structures and allele group frequencies, thereby broadening our understanding of the *KIR* gene complex.

**Keywords: KIR, HSCT, haplotype frequency, donor selection, immunogenetics**

## INTRODUCTION

The Killer-cell immunoglobulin-like receptor (*KIR*) gene cluster is located on the long arm of human chromosome 19 (19q13.4). The family includes the 15 expressed genes *KIR2DL1-4*, *KIR2DL5A*, *KIR2DL5B*, *KIR3DL1-3*, *KIR2DS1-5*, and *KIR3DS1* and the two pseudogenes *KIR2DP1* and *KIR3DP1* (1). *KIR* genes are characterized by a high level of sequence homology among different gene loci on one hand, and multi-layer complexity at the gene and haplotype level on the other hand (2, 3). The *KIR* gene copy number (CN) of a locus varies from 0 to 3 per haplotype (4). Each of the *KIR* loci displays high allelic polymorphism. One thousand one hundred ten allele variants and 543 KIR proteins have been named so far [Immuno Polymorphism Database IPD-*KIR* v2.9.0, December 2019 (5)]. In addition, diversity of the KIR gene region was found to be shaped in a population-specific manner through evolutionary mechanisms (6). Frequent occurrence of hybrid *KIR* genes (7–14), as well as events of alternative splicing of *KIR* genes, resulting in potentially functional receptor isoforms (15), have been described.

The high genetic complexity may be rooted in the diverse functions of *KIR* gene products in the human innate immune system. KIR proteins are transmembrane receptors primarily expressed by Natural Killer (NK) cells. Most KIR are located in the NK cell plasma or endosomal membrane and are known to be involved in the regulation of the immune response to viral infections and cancer or in the governance of histocompatibility during pregnancy (16–18). Bound to their corresponding ligands, which include classical and non-classical human leukocyte antigen (HLA) class I molecules, they convey activating or inhibitory signals to the NK cell (18, 19). *KIR3DL1-3, KIR2DL1-3*, and *KIR2DL5A/B* are the KIR receptors with inhibitory potential, the activating receptors are KIR3DS1 and KIR2DS1-5. The integrated signals of KIR and further NK cell receptors regulate NK cell activity between the two extremes of tolerance and killing (16). *KIR2DL4* plays a special role in the *KIR* gene family. This receptor appears to participate in the NK-mediated control of the maternal/fetal interface during pregnancy and is probably not involved in the cancer or infection surveillance (20, 21). However, lack of the *KIR2DL4* in maternal NK cells is still compatible with successful pregnancy (22).

In present-day unrelated hematopoietic stem cell transplantation (HSCT) practice, the alleles of particular *HLA* genes are matched between patient and stem cell donor to avoid undesired immune reactions of donor T-cells against the host organism (23–25). However, as graft-versus-host disease (GvHD) and disease relapse remain serious complications for many patients, there is an ongoing quest to identify further immunogenic factors with the potential to improve the outcome of HSCT. Since NK cells are the first lymphocytes to reconstitute in patients after HSCT (26), efforts have been made to exploit their potential to elicit a rapid and targeted immune response against remaining leukemic cells (27–32). Recent research indicates an impact of donor *KIR* genotype on the outcome of HSCT, but results remain controversial (33–37).

The knowledge of population-specific allelic haplotype frequencies for *HLA* genes of the major histocompatibility complex (MHC) offered advantages in the field of unrelated donor HSCT. For instance, all major donor search algorithms utilize *HLA* haplotype frequencies to estimate the probabilities for listed donors with ambiguous or incomplete *HLA* typing data to be a match for a specific patient (38, 39). Matching probability analyses using population-specific *HLA* haplotype frequencies allow targeted planning of donor center recruitment strategies (40). Beyond that, a positive impact on outcome of HSCT is proposed for the additional matching of non-*HLA* genes located in the MHC region via haplotype matching between patient and donor, but study results are still inconclusive (41–43).

Efficient use of *KIR* genes in donor selection for successful HSCT accordingly would benefit from knowledge of population-specific *KIR* haplotype frequencies. However, compared to *HLA*, *KIR* haplotype inference from separately typed loci is much more difficult because of the high number of possible diplotypes one unphased *KIR* genotype is compatible with. For example, two alleles genotyped at one *KIR* locus may be located either both on the same chromosome or each alone on one of the two chromosomes. These ambiguities and the high number of *KIR* loci lead to a large amount of possible allelic diplotypes that requires unrealistically large memory for the implementation of a conventional expectation-maximization (EM) algorithm.

A way to cope with the high number of possible allelic diplotypes per individual and to still reproduce the polymorphic nature of *KIR* haplotypes is to limit the underlying haplotype structure to pre-established patterns (44, 45). Previous research analyzed the haplotype structure of the *KIR* gene complex at various levels of resolution. Presence or absence of certain *KIR* loci in individuals was found not to occur at random, but to follow structural patterns. At the level of gene content (presence/absence, P/A) polymorphism, two major groups of *KIR* haplotypes (A and B) are described (2, 19). A and B haplotypes are characterized and distinguished by the presence of specific sets of *KIR* genes alongside the four "framework" genes (*KIR3DL3*, *KIR3DP1*, *KIR2DL4*, and *KIR3DL2*), flanking the centromeric and telomeric regions of the vast majority of all *KIR* haplotypes. *KIR2DS4* is the only activating receptor encoded by A haplotypes, whereas B haplotypes carry up to 5 activating KIR. More detailed analyses of *KIR* haplotypes and their frequencies in different cohorts were undertaken at gene content level (45–48), copy number level (4, 14, 45, 49, 50) and at (partial) allelic resolution level (44, 48, 49, 51, 52). All studies confirmed the basic concept of A/B haplotypes, but also documented numerous deviations from these structural patterns, caused, e.g., by recombination, gene fusion, deletion, or insertion events.

In our study, we extended the method to limit haplotype structure to pre-established patterns in order to make it applicable to our data obtained by high-throughput *KIR* genotyping in the context of unrelated HSCT donor registration (53, 54). We applied a three-step approach to estimate population-specific *KIR* haplotype frequencies at allele group resolution. We analyzed the *KIR* genotypes in a cohort of $n = 458$ families in order to reduce phasing and typing ambiguities and thus the number of possible diplotypes for the $n = 916$ parents. Structural complexity was further confined by restricting possible haplotypes to a set of 92 previously described *KIR* copy number haplotypes. Haplotype frequencies were then derived using an implementation of the EM algorithm that dealt with remaining ambiguities.

# MATERIALS AND METHODS

## *KIR* and *HLA* Genotyping

Between October 2016 and April 2019, 2.6 million donors recruited by DKMS were *KIR* genotyped at allelic resolution by DKMS Life Science Lab in Dresden, Germany, using next generation sequencing methods (53, 55, 56). Genotyping comprises the *HLA* loci *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1*, and *-DPB1*, the *KIR* loci *KIR2DL1-5*, *KIR3DL1-3*, *KIR2DS1-5*,

**Abbreviations:** CNPS, copy number haplotype pattern set; EM, expectation-maximization; GvHD, graft-versus-host disease; HF, haplotype frequency; HLA, human leukocyte antigen; HSCT, hematopoietic stem cell transplantation; KIR, Killer-cell immunoglobulin-like receptor; LD, linkage disequilibrium; MHC, major histocompatibility complex; NK, natural killer.

*KIR3DS1*, *KIR2DP1*, and *KIR3DP1*, as well as *ABO* (57), *RhD*, *CCR5* (58), and *MIC-A/B* (59). The KIR genotyping approach delivers both allele group information and copy numbers for every KIR gene (53). DNA was extracted from blood samples or buccal swabs with the informed consent of the donors.

For the analysis of *KIR* haplotypes, *KIR* genotyping results were shortened to the first three digits of the allele name, thereby merging alleles with synonymous substitutions within the coding region and non-coding mutations. Allelic ambiguities due to variations outside the typed exons (exons 3, 4, 5, 7, 8, and 9) were grouped and denoted by a trailing "c" (**Supplementary Information S1**). In the following, 3-digit allelic level and "c"-groups are referred to as "allele group" resolution. Since a clear distinction of genes of the loci *KIR2DL5A* and *KIR2DL5B* was not possible without sequence information on exon 1 and promoter region, we treated *KIR2DL5A* and *KIR2DL5B* as one locus. We considered 16 *KIR* loci in total.

## Family Selection

Information on family relations is not recorded during DKMS donor recruitment. Data retrieval from the DKMS Germany donor file, demanding consistency of addresses and surnames and a minimal age difference of >20 years between parents and offspring, yielded a pseudonymized set of potential families of self-assessed German origin. The genetic relationship between potential family members was verified on the basis of the respective *HLA-A*, *-B*, *-C*, *-DRB1*, and *-DQB1* typing data. A cohort of 458 *HLA*-confirmed families with two parents and at least one child were included into our study. Four hundred two of the families had one child, 55 had two, and one family had three children registered with the donor center ($n_{total} = 1{,}431$).

## *KIR* Data Refinement via Family Information

*KIR* genes of all members of our family cohort were resolved to allele group level and copy number. Knowledge of family relations was used to review *KIR* typing when comparison of data between parents and children revealed ambiguities or inconsistencies. For instance, in 1.6% of the 14,656 typed parental *KIR* loci (916 samples × 16 loci), the allele group typing results of a child apparently did not match the parental genotypes. All but 13 cases, concerning in total 10 sets of parents, could be solved by re-inspection of the sequencing data and application of the family information. The matching status of new and so far unnamed alleles between the family members (464 cases) was verified by comparison of sequencing data and was accordingly considered during haplotype inference. Families with more than one child ($n = 56$) permitted investigation of potential recombination events by intersecting parental haplotypes deduced via *KIR* genotyping data of different children.

## Copy Number Haplotype Pattern Set (CNPS)

In order to reduce haplotype complexity originating from phasing ambiguities between *KIR* loci, only *KIR* genotypes that could be split into two haplotypes that met the structural pattern of 92 previously described *KIR* copy number haplotypes were permitted. The set of 92 *KIR* copy number haplotypes used as reference pattern is hereinafter referred to as CNPS. The CNPS included 12 copy number haplotypes described by Pyo et al. (48), 52 by Jiang et al. (4), 27 by Pyo et al. (50), and one by Roe et al. (14) (**Supplementary Information S2**). Because nomenclature of *KIR* haplotypes in the different publications in not consistent, we assigned a code name (column "HT code" in **Supplementary Information S2**) to every copy number haplotype in the CNPS which will be used below.

*KIR* loci *KIR2DL5, KIR2DS3*, and *KIR2DS5* have experienced a duplication event in the past and can be found in both the centromeric and the telomeric section of particular haplotypes of the *KIR* gene complex (60, 61). We could not distinguish between the respective centromeric and telomeric variants of the three loci in our analysis. For copy number haplotypes from publications where *KIR2DS3* and *KIR2DS5* were treated as one single locus and where thus the assignment to one of the loci was ambiguous, we split the haplotype into all possible copy number forms. For example, a haplotype which specifies 2 copies of locus *KIR2DS3S5* was split into three allowed copy number haplotypes: one haplotype with one copy of *KIR2DS3* and *KIR2DS5*, each, and two haplotypes with two copies of one of the two genes and none of the other. Copy number haplotypes marked as containing hybrid or fusion gene loci by Pyo et al. (50) and Roe et al. (14) were disregarded.

## Workflow

All software required for *KIR* haplotype frequency (HF) estimation in our approach was written in Perl 5. A schematic overview of the algorithm is shown in **Figure 1**. Families were analyzed in mother-father-child sets, i.e., a family with two children was analyzed in two separate family sets. Only diplotypes of the 916 parents were allowed to pass into the subsequent *KIR* HF estimation. Parents whose haplotypes had already counted for frequency estimation via a first family set were flagged to avoid duplication. For each of the mother-father-child constellations of our data set, all possible allele group diplotypes of mother and father were deduced. This was done by locus-wise diplotype inference and subsequent iteration of all possible allele constellations of all loci. The extensive diplotype list was then filtered with the CNPS to exclude all diplotypes that could not be explained by a pair of the allowed copy number haplotypes. In order to limit artifacts from lower-resolution typing data, we restricted the number of possible diplotypes per individual to a maximum of $n = 1{,}000{,}000$. Parents with more possible diplotypes after application of the CNPS were excluded from the subsequent HF estimation, unless they passed this requirement in a family constellation with another child. The most likely set of *KIR* HFs was finally derived from the parental diplotypes by means of an expectation-maximization (EM) algorithm (62). Haplotypes were initialized with equal frequencies prior to the start of the EM algorithm. The stop criterion, which defines the allowed maximal HF change between consecutive estimations, was set to $5^{*}10^{-5}$. HFs were cut after a minimal frequency of $f = 1/2n$, with $n$ being the number of individuals in the final haplotype estimation, corresponding to the occurrence of at least one

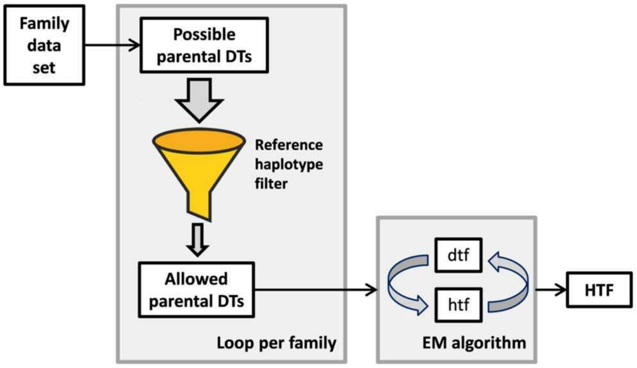


**FIGURE 1 |** Simplified workflow of our haplotype frequency estimation approach. The number of 16-locus *KIR* diplotypes per individual is reduced by inference from a family context and subsequent filtering via a set of reference copy number haplotypes. Only diplotypes that can be explained by a pair of reference haplotypes are allowed. Remaining ambiguities are resolved in a conventional expectation-maximization (EM) algorithm. DT, diplotype; dtf, preliminary diplotype frequencies; htf, preliminary haplotype frequencies; HTF, final haplotype frequencies.

haplotype in the sample. The resulting allele group haplotypes were attributed to the corresponding CNPS haplotypes in order to analyze the allelic diversity within the different copy number structures.

HFs were estimated both for the entire *KIR* gene region and separately for the centromeric and telomeric sections. In order to be able to assign the allele combinations of gene cluster *KIR2DL5*∼*KIR2DS3*∼*KIR2DS5* to a centromeric or telomeric location, we included these genes in both estimations of partial *KIR* HFs.

The approach to reduce the ambiguous nature of the *KIR* gene complex by means of the CNPS also permits the imputation of allelic haplotypes from large cohorts of individuals without a family context. In order to investigate the additional benefit of phase resolution via family affiliation, allele group diplotypes of all 916 parents of our cohort were determined in a separate workflow without the phasing knowledge of their families. Filtering of the diplotypes via the CNPS and estimation of the allele group *KIR* HF by the EM algorithm was carried out according to the family workflow described above.

## Linkage Disequilibrium

The linkage disequilibrium (LD) coefficient D' was calculated from the entire HF set (before cut below $f = 1/2n$) for any two-locus combination (63, 64). Significance was tested with Fisher's exact test. *P*-values were subjected to a Holm-Bonferroni correction for multiple testing. Multi-locus LD analysis was conducted using the normalized entropy difference ε between the observed HFs and those expected under the null hypothesis of linkage equilibrium (64, 65). The value of ε ranges between 0 and 1, with larger values indicating stronger LD. An absent locus ("NEG") was treated as one allele variant.

# RESULTS

## *KIR* Genotyping Resolution of the Family Cohort

91.8% of the 22,896 (1,431 donors × 16 loci) typed loci in our final *KIR* family data set were determined to allelic resolution, absence, or allele group level. Of this share, 74.7% of the identified copy number equivalents were resolved to allelic level, 21.5% to absence of the locus and 3.8% to broader allele groups, where remaining ambiguities impeded the differentiation between two or more alleles.

The remaining 8.2% of the 22,896 typed loci carried ambiguities either due to unknown phasing of the typed exons (e.g., *KIR2DL1*003c* + *KIR2DL1*004*|*KIR2DL1*006* + *KIR2DL1*010*) or to uncertain copy number determination (e.g., *KIR2DL3*001* + *KIR2DL3*002*| *KIR2DL3*001* + *KIR2DL3*001* + *KIR2DL3*002*), allowing for multiple valid allele (group) pairs.

## Parental Cohort

In the final analysis, *KIR* data of 790 of the 916 parents was used to estimate HFs of the entire *KIR* gene region. The 790 individuals originate from 403 of the 458 families. Both parents were included into the HF estimation for 387 families, 16 further families contributed one parent, each.

Reasons for exclusion from the final analysis set were: 97 parents (10.6%) were excluded because no combination of the allowed 92 copy number haplotypes could explain the deduced diplotypes. Twenty parents (2.2%) were excluded because of mismatching to all of their potential children in one *KIR* locus (nine families) or two *KIR* loci (one family). Seven further parents (0.8%) were excluded because the number of possible diplotypes after the filtering step exceeded the threshold ($n$ = 1,000,000) of possible *KIR* diplotypes per individual. Finally, both parents (0.2%) of one family were omitted because the respective diplotype estimations on the basis of the *KIR* data of two children showed no intersection, but indicated a possible recombination event in the father's chromosomes (**Supplementary Information S3**). In this case, comparison of the father's different haplotype sets derived from two children revealed an exchange in alleles in locus *KIR3DL3*.

Consideration of only centromeric or telomeric genes for *KIR* haplotype estimation in the family context allowed the inclusion of more parents. HFs were calculated from KIR data of 838 (876) individuals for the centromeric (telomeric) loci. The number of individuals excluded because no combination of valid CNPS haplotypes could explain the diplotypes decreased to 6.6% (3.5%). Omission because of mismatching was reduced to 1.7% (0.7%), None of the parents were excluded due to high numbers of possible diplotypes.

When we deduced allele group diplotypes of the 916 parents for the entire *KIR* gene region without the phasing knowledge of the family relations, the parental cohort that could be included was considerably smaller. Applying the same filtering and configuration values as in the family approach, *KIR* data of only 438 (47.8%, compared to 790 or 86.2% in the family calculations) individuals passed into the HF estimation. On the one hand, only 6 individuals (0.7% vs. 10.6% in the family approach) were omitted because no valid combination of the reference copy number haplotypes could explain their diplotypes. On the other hand, however, the number of possible diplotypes exceeded the threshold of $n = 1{,}000{,}000$ in 473 parents (51.6 vs. 0.8% in the family approach), demonstrating the efficiency of ambiguity reduction via family information.

## *KIR* Haplotype Frequencies for the Entire Gene Region

*KIR* HF estimation for the data of the 790 parents resulted in 551 different *KIR* allele group haplotypes with $f \geq 1/2n$, corresponding to a frequency sum of 90.8% (**Figure 2**, **Supplementary Information S4**). The 20 most frequent allele group *KIR* haplotypes are listed in **Table 1** and comprise a cumulated frequency of 26.0%.

Of the 92 permitted reference copy number haplotypes, only 21 are represented in the 551 estimated allele group haplotypes. **Table 2** shows identity and frequencies of these CNPS haplotypes. Where the assignment was possible, the A/B haplotype nomenclature established by 48 is indicated. The most frequent of the CNPS haplotypes is P10_01 (cA01~tA01) with $f = 59.6\%$.

## Allelic Diversity

We attributed the allele group haplotypes to the respective copy number haplotypes in order to analyze the allelic diversity within the different copy number structures (**Supplementary Information S5**). However, due to the different frequencies of the copy number haplotypes and thus the different chances to detect allelic diversity, observations in our dataset can only be seen as indications of patterns. For the analysis, we considered only eight copy number haplotypes with a frequency of $f \geq 1\%$ (**Table 2**). The number of allele group haplotypes per copy number haplotype varies between 8 and 318, with a clear positive correlation between copy number HF and the number of allele group haplotypes. All eight copy number haplotypes are combinations of only three centromeric (cA01, cB01, and cB02) and two telomeric (tA01 and tB01) motifs (in the nomenclature introduced by Pyo et al. (48).

**Figure 3** shows for each KIR gene the allele frequencies overall and for each copy number haplotype. Marked differences in overall allelic variability can be observed between the different genes with the highest diversities in the two outermost framework genes *KIR3DL3* ($n = 29$ named alleles) and *KIR3DL2* ($n = 17$), as well as in *KIR3DL1* ($n = 14$). Allelic variability and identity within one locus, however, depends clearly on the respective haplotype motif. Essentially, A haplotype motifs have a higher allelic diversity than B haplotype motifs in both, the centromeric and the telomeric *KIR* section. In addition, the different structural motifs also correlate with different alleles. For example, in locus *KIR3DL2*, tB01 haplotypes (P10_02, _03, _06, _07, and _10) are clearly dominated by allele *007 with allele frequencies between 65.0 and 100%. tA01 haplotypes (P10_01, _04, and _08), in contrast, have a much higher allelic variability in *KIR3DL2* where allele *007 only reaches a maximum frequency of 0.4%. Similar patterns can be seen in all framework genes and in genes *KIR2DP1* and *KIR2DL1*, which are present in both, A and B haplotypes.

The allocation of allele group haplotypes to the copy number haplotypes also reveals conserved allele combinations. For instance, the partial haplotype *KIR2DL2~KIR2DL3~KIR2DL1* that exhibited linkage disequilibrium also in the LD analysis (see below) displays an interesting pattern of allelic distribution. cB1 haplotypes (P10_04, _06, and _07) are almost exclusively populated with allelic haplotype block 001~NEG~004, while cA01 haplotypes (P10_01, _02, and _03) are dominated by allele group combinations NEG~001~003c and NEG~002~001c. Allele *003 prevails in cB02 haplotypes (P10_08 and _10) with only *KIR2DL2* present. A second example is the gene cluster *KIR2DL5~KIR2DS3~KIR2DS5*. Our typing method does not allow the unambiguous assignment of this gene cluster to a centromeric or telomeric localization. However, the allele distribution of these genes in the different copy number haplotypes clearly reflects a conserved linkage of certain allele

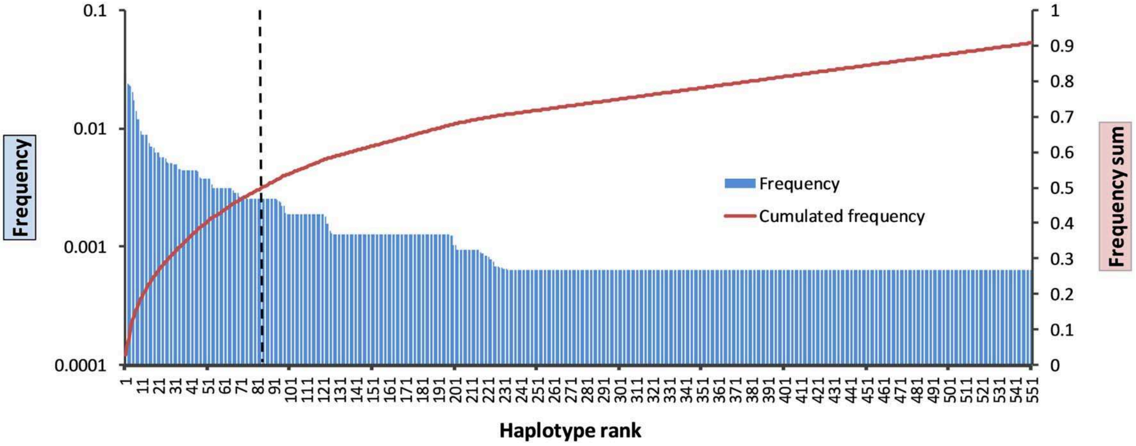


**FIGURE 2 |** Frequency distribution of the 551 allele group *KIR* haplotypes above the resolution threshold of $f = 1/2n$. Blue bars: haplotype frequencies; red line: cumulated frequency; dashed black line: haplotype rank 84 with 50% cumulated haplotype frequency.

**TABLE 1 |** Top 20 allele group *KIR* haplotypes ranked by their respective frequencies.

| A/B nomenclature | HT code | 3DL3 | 2DS2 | 2DL2 | 2DL3 | 2DP1 | 2DL1 | 3DP1 | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5 | 2DS3 | 2DS5 | 2DS1 | 3DL2 | Frequency |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cB02~tA01 | P10_08 | **003** | **001c** | **003** | NEG | NEG | NEG | **001c** | **001** | **002** | **001c** | NEG | NEG | NEG | NEG | NEG | **002c** | 0.029046 |
| cA01~tA01 | P10_01 | **001** | NEG | NEG | **002** | **003** | **001c** | **006** | **008** | **004** | **006** | NEG | NEG | NEG | NEG | NEG | **005** | 0.023675 |
| cA01~tA01 | P10_01 | **002** | NEG | NEG | **001** | **NEW** | **003c** | **001c** | **011** | **005** | **010** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.023363 |
| cA01~tA01 | P10_01 | **001** | NEG | NEG | **001** | **002** | **003c** | **001c** | **001** | **015c** | **001c** | NEG | NEG | NEG | NEG | NEG | **002c** | 0.022692 |
| cA01~tA01 | P10_01 | **009** | NEG | NEG | **001** | **NEW** | **003c** | **001c** | **011** | **005** | **010** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.020308 |
| cA01~tA01 | P10_01 | **008** | NEG | NEG | **001** | **005** | **003c** | **001c** | **008** | **001c** | **003** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.017089 |
| cB01~tB01 | P10_06 | **003** | **001c** | **001** | NEG | **001** | **004** | **001c** | **005** | NEG | NEG | **013c** | **002c+002c** | **001c+002** | NEG | **002c** | **007** | 0.014089 |
| cA01~tA01 | P10_01 | **001** | NEG | NEG | **002** | **003** | **001c** | **NEW** | **001** | **002** | **001c** | NEG | NEG | NEG | NEG | NEG | **002c** | 0.012025 |
| cA01~tA01 | P10_01 | **009** | NEG | NEG | **001** | **002** | **003c** | **001c** | **011** | **005** | **010** | NEG | NEG | NEG | NEG | NEG | **010** | 0.012025 |
| cA01~tA01 | P10_01 | **013** | NEG | NEG | **002** | **003** | **001c** | **006** | **008** | **004** | **006** | NEG | NEG | NEG | NEG | NEG | **009** | 0.009494 |
| cA01~tA01 | P10_01 | **013** | NEG | NEG | **002** | **003** | **001c** | **NEW** | **006** | **007c** | **004** | NEG | NEG | NEG | NEG | NEG | **008** | 0.008861 |
| cA01~tA01 | P10_01 | **019** | NEG | NEG | **002** | **003** | **001c** | **006** | **008** | **004** | **006** | NEG | NEG | NEG | NEG | NEG | **005** | 0.008861 |
| cA01~tA01 | P10_01 | **001** | NEG | NEG | **002** | **003** | **001c** | **NEW** | **008** | **001c** | **003** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.008861 |
| cA01~tA01 | P10_01 | **002** | NEG | NEG | **001** | **002** | **003c** | **001c** | **008** | **001c** | **003** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.008769 |
| cA01~tB01 | P10_03 | **001** | NEG | NEG | **001** | **002** | **003c** | **001c** | **005** | NEG | NEG | **013c** | **001c** | NEG | **002** | **002c** | **007** | 0.007493 |
| cA01~tA01 | P10_01 | **001** | NEG | NEG | **002** | **003** | **001c** | **006** | **011** | **005** | **010** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.007104 |
| cA01~tA01 | P10_01 | **002** | NEG | NEG | **001** | **005** | **003c** | **001c** | **008** | **001c** | **003** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.006962 |
| cA01~tB01 | P10_03 | **009** | NEG | NEG | **001** | **002** | **003c** | **001c** | **005** | NEG | NEG | **013c** | **001c** | NEG | **002** | **002c** | **007** | 0.006842 |
| cB02~tA01 | P10_08 | **003** | **001c** | **003** | NEG | NEG | NEG | **001c** | **008** | **001c** | **003** | NEG | NEG | NEG | NEG | NEG | **001c** | 0.006329 |
| cA01~tA01 | P10_01 | **001** | NEG | NEG | **002** | **003** | **001c** | **005** | **001** | **002** | **001c** | NEG | NEG | NEG | NEG | NEG | **002c** | 0.006329 |

*Framework genes are shaded in gray. Alleles of present loci are indicated in bold letters. A/B haplotype nomenclature according to Pyo et al. (48). HT code, code of the copy number haplotype pattern (see* ***Supplementary Information S2****); c, centromeric; t, telomeric; ~, point of separation between framework genes KIR3DP1 and KIR2DL4.*

group combinations that reveal the position of the cluster in the *KIR* gene region. On the one hand, P10_04 (cB01~tA01), representing a centromeric localization of the cluster, is almost exclusively composed of *KIR2DL5**002c~*KIR2DS3**001c. A very small fraction of *KIR2DS3* alleles is "NEW". On the other hand, tB01, the common haplotype structure of P10_02, P10_03, and P10_10, is known to include a telomeric *KIR2DL5*~*KIR2DS3*~*KIR2DS5* gene cluster. While the exclusive allelic composition in CN haplotypes P10_03 and P10_10 in our data is 2DL5*001c~2DS5*002, P10_02 carries either of the two combinations 2DL5*002c~2DS3*002 and 2DL5*002c~2DS3*003N. The conserved centromeric and telomeric allele combinations in our data set are listed in the **Supplementary Information S6**.

## Linkage Disequilibrium

LD coefficient D' was calculated for all two-locus haplotypes. Altogether, 368 pairs with significant deviation from equilibrium were identified (**Supplementary Information S7**), 129 of them with positive D'. In a large number of the significant cases of LD, one or both of the "alleles" is the absent locus. Thirty-two pairs of actual alleles in significant LD with $D' \geq 0.9$ and a HF of $f \geq 0.1$ are listed in **Table 3**. Eleven of the pairs map to the centromeric section of the *KIR* gene complex, 18 to the telomeric section. The three pairs including loci *KIR2DL5* and *KIR2DS5* show close linkage to *KIR2DL4* and thus are probably located on the telomeric part of the *KIR* gene cluster.

**Figure 4** shows the LD between the 16 analyzed loci of the *KIR* complex. The normalized entropy difference ε indicates a linkage disequilibrium within the designated centromeric and telomeric parts of the *KIR* gene complex, but much less between the two parts. For the pair of the two inner framework genes *KIR3DP1* and *KIR2DL4*, the value is close to equilibrium ($\varepsilon = 0.07$). The highest telomeric LD ($\varepsilon = 0.42$) is found for the loci *KIR3DL1* and *KIR2DS4*. In the vast majority of haplotypes, amongst others in the by far most frequent copy number haplotype P10_01(cA01~tA01), those two loci are adjacent and either both present or both absent. In addition, *KIR3DL1* is in LD with framework gene *KIR2DL4* ($\varepsilon = 0.38$), and *KIR2DL4* with *KIR2DS4* ($\varepsilon = 0.39$). In the centromeric stretch, the highest LD ($\varepsilon = 0.37$) is found for *KIR2DS2* and *KIR2DL2*, two adjacent and usually concurrent loci. A further block of linkage disequilibrium is formed by loci *KIR2DL3*, *KIR2DP1*, and *KIR2DL1* ($0.32 \leq \varepsilon \leq 0.36$). The three loci *KIR2DL5*, *KIR2DS3*, and *KIR2DS5* play a special role in the multi-locus LD analysis since our typing method does not allow the separation of *KIR2DL5A* and *KIR2DL5B* and thus the direct assignment of the *KIR2DL5*~*KIR2DS3*~*KIR2DS5* cluster to centromeric or telomeric position on the chromosome. The linkage analysis reveals LD for the two pairs *KIR2DL5*~*KIR2DS3* ($\varepsilon = 0.32$) and *KIR2DL5*~*KIR2DS5* ($\varepsilon = 0.30$), while the two loci *KIR2DS3* and *KIR2DS5* are almost in linkage equilibrium ($\varepsilon = 0.01$). Overall, the linkage of the three genes of this cluster is higher to loci of the telomeric than to loci of the centromeric region. *KIR2DS5* shows no linkage to genes of the centromeric *KIR* region at all. This is consistent with the conserved centromeric and telomeric allele combinations observed for the *KIR2DL5*~*KIR2DS3*~*KIR2DS5* cluster, according to which *KIR2DS5* is not present in centromeric position.

**TABLE 2 |** List of the 21 *KIR* copy number haplotypes represented in the 551 allele group *KIR* haplotypes above the resolution threshold and their frequencies.

| 3DL3 | 2DS2 | 2DL2 | 2DL3 | 2DP1 | 2DL1 | 3DP1 | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5 | 2DS3 | 2DS5 | 2DS1 | 3DL2 | Frequency | HT code | A/B nomenclature |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.59554 | P10_01 | cA01~tA01 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.09810 | P10_08 | cB02~tA01 |
| 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0.07603 | P10_03 | cA01~tB01 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0.05190 | P10_04 | cB01~tA01 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 1 | 1 | 0.02870 | P10_06 | cB01~tB01 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | 1 | 0.01518 | P10_07 | cB01~tB01 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0.01266 | P10_10 | cB02~tB01 |
| 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0.01245 | P10_02 | cA01~tB01 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00253 | P10_09 | cB02~tB01 |
| 1 | 1 | 1 | 0 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.00253 | J12_13 | cB$^{*}$~t$^{\#}$ |
| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.00253 | J12_22 | cA$^{*}$~tB$^{*}$ |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0.00190 | J12_12 | cB$^{*}$~tB$^{*}$ |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.00127 | J12_27 | cA$^{*}$~tA1 |
| 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.00127 | J12_53 | cB$^{*}$~tA1 |
| 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.00127 | J12_65 | cA$^{*}$~tA1 |
| 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.00063 | P12_01 | cA01~tA01-del5 |
| 1 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 0 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 0.00063 | J12_28 | cA$^{*}$~tB$^{*}$ |
| 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.00063 | P12_04 | cB01\|tA01-del9 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.00063 | J12_60 | cA$^{*}$~tA1 |
| 1 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0.00063 | J12_16 | cA$^{*}$~t$^{\#}$ |
| 1 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0.00063 | J12_30 | cB02\|tA01-ins4 |

*Framework genes are shaded in gray. A/B haplotype nomenclature according to Pyo et al. (48, 50) was added where the assignment was possible. HT code, code of the copy number haplotype pattern; c, centromeric; t, telomeric; ~, point of separation between framework genes KIR3DP1 and KIR2DL4;* $^{*}$*, affiliation to a general A or B structure but deviation from the established nomenclature;* $^{\#}$*, distinction between A and B not possible.*

## Assignment of *KIR2DL5~KIR2DS3~KIR2DS5* to Centromeric or Telomeric Position on the Haplotype

The analysis of the allelic diversity of different copy number haplotype structures revealed the association of conserved allele combinations of gene cluster *KIR2DL5~KIR2DS3~KIR2DS5* with its localization on the chromosome (**Supplementary Information S6**). We applied this information to assign centromeric and telomeric positions of the gene cluster in our KIR HFs.

Of the 551 *KIR* allele group haplotypes, 164 contained the respective genes in a total of 11 different combinations. In 96.3% of the cases, the localization of the *KIR2DL5~KIR2DS3~KIR2DS5* genes could be assigned unambiguously. Six cases in two allele combinations included "NEW" alleles in *KIR2DS3*. For these cases we deduced the position of the clusters in the context of the overall haplotype structure. KIR HFs for the entire gene region with inferred localization of gene clusters *KIR2DL5~KIR2DS3~KIR2DS5* are shown in the **Supplementary Information S8**.

## Frequencies of Partial *KIR* Haplotypes

Estimation of *KIR* HFs of the centromeric gene region from *KIR* data of 838 parents resulted in 323 different allele group haplotypes with $f \geq 1/2n$ (frequency sum 97.1%), representing 22 partial copy number haplotypes (**Supplementary Information S9**). Frequency of the three structural haplotypes cA01, cB01, and cB02 reached a total of 92.8%. Assignment of cluster *KIR2DL5~KIR2DS3~KIR2DS* to the centromeric haplotypes was deduced as described above. Interestingly, the partial HFs of the centromeric *KIR* genes reveal an additional allele combination of the *KIR2DL5~KIR2DS3~KIR2DS5* cluster. One haplotype with the combination *KIR2DL5**019B~ *KIR2DS5**008 was found, the only centromeric occurrence of *KIR2DS5* in our study. Its frequency ($f = 5.97^{*}10^{-4}$) indicates a unique occurrence of the haplotype in our sample.

Accordingly, telomeric KIR HFs, including *KIR* data of 876 parents yielded 146 allele group haplotypes with $f \geq 1/2n$ (frequency sum 98.8%). Only 12 partial copy number haplotypes were represented. The frequency of copy number haplotypes tA01, tB01 (KIR2DS3 present), and tB01 (KIR2DS5 present) added up to 96.8% (**Supplementary Information S10**).

## DISCUSSION

We present allele group *KIR* haplotypes estimated from 790 parents of a cohort of $n = 403$ German families. The restriction of haplotype structures to a set of copy number haplotypes from previous publications enabled us to cope with the enormous structural ambiguities of the *KIR* gene complex. Deducing diplotypes within the context of families allowed us to further reduce the number of possible diplotypes per individual. Our

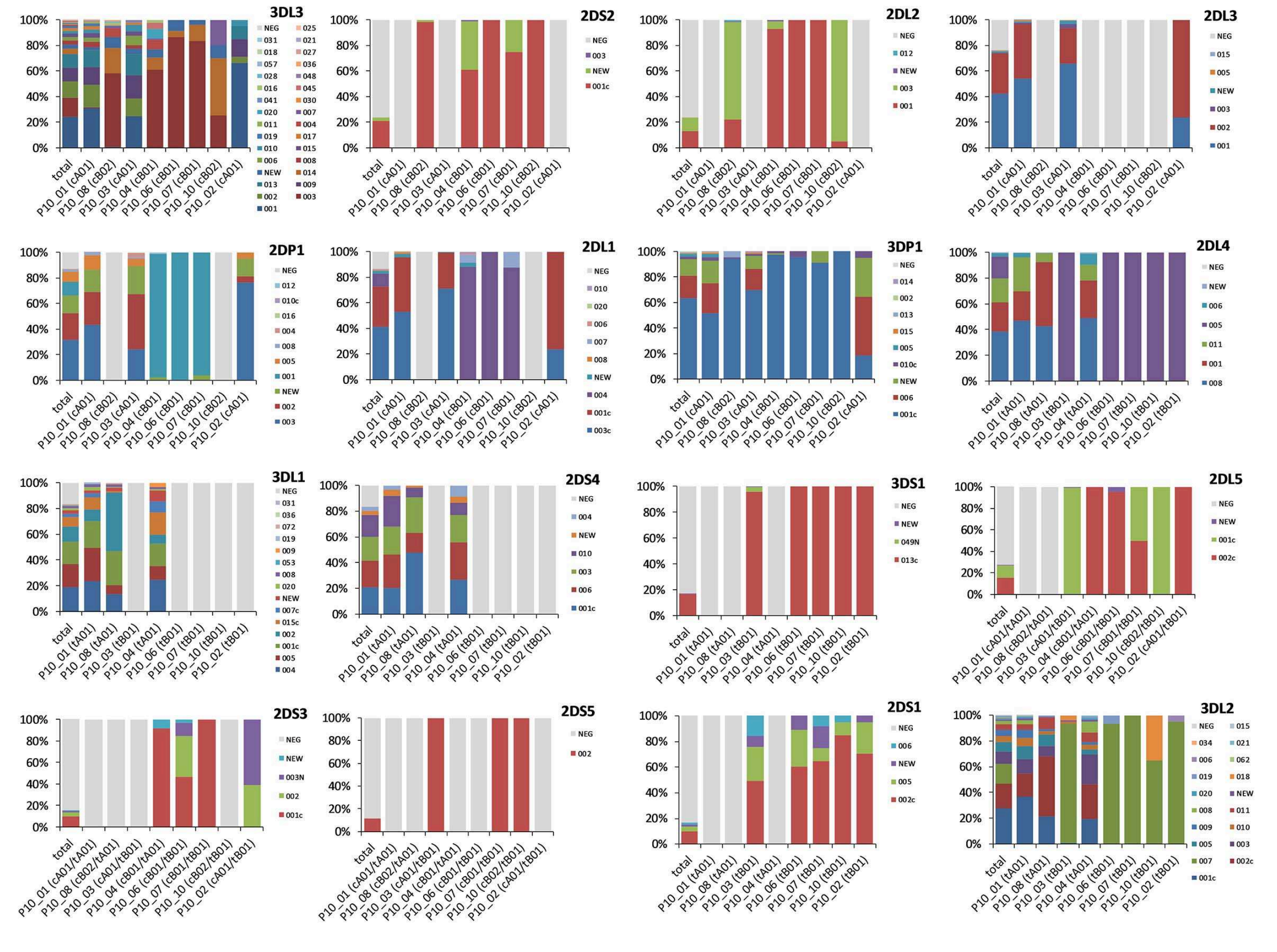


**FIGURE 3 |** Allele group frequencies per *KIR* gene. Displayed are, for each of the 16 *KIR* genes, the allele frequencies for the estimated allele group KIR HF overall (total) and for the 8 CNPS haplotypes with a frequency of $f \geq 1\%$. The proportion of absence of the respective gene is indicated as "NEG" in light gray. The centromeric or telomeric haplotype structure of the respective CNPS haplotypes in A/B haplotype nomenclature according to Pyo et al. (48, 50) is indicated in brackets in the labeling of the diagram axes.

approach yielded a set of 551 allele group *KIR* haplotypes, representing 21 of the 92 CNPS haplotypes.

The results of our study are in good accordance with two other studies on allelic *KIR* haplotypes. A table comparing our Top20 haplotypes with findings of Vierra-Green et al. (44), Hou et al. (52) can be found in the **Supplementary Information S11**. Both former studies inferred their allelic HFs from cohorts of European ancestry without a family context and did not include information on the polymorphism of the pseudogene loci *KIR2DP1* and *KIR3DP1* in their haplotype determination. Haplotype structures were restricted to smaller reference sets of content haplotypes.

Nineteen of our Top 20 allele group *KIR* haplotypes correspond to allelic haplotypes published by Vierra-Green et al. (44). For the most frequent haplotype in both studies, *KIR3DL3* 003~KIR2DS2*001c~KIR2DL2*003~KIR3DP1*001c ~KIR2DL4* 001 ~ KIR3DL1* 002 ~ KIR2DS4* 001c ~ KIR3DL2* 002c*, we estimated a frequency of $f = 0.0290$, while Vierra-Green et al. determined $f = 0.0306$. During comparison of the allelic haplotypes it became apparent that several haplotypes in Vierra-Green et al. carried allele *KIR2DS4*007* which never occurred in our haplotype set. Analysis of the sequence provided a potential explanation: *KIR2DS4*007* differs from *KIR2DS4*010* only in the first base of exon 4. As *KIR2DS4*010* was not listed in the IPD-*KIR* database before version 2.2.0 (May 2010), this allele may not have been included in allele typing and interpretation of the former study. Our typing routine distinguishes between the two alleles and we find almost exclusively *KIR2DS4*010* in our donors.

In a study published by Hou et al. (52), the centromeric and telomeric stretches of the *KIR* gene complex were analyzed separately and allelic haplotypes were grouped into different consensus structures. Our estimated allele group *KIR* haplotypes are in good agreement with their findings (**Supplementary Information S7**). The centromeric parts of 19 of our Top 20 allele group haplotypes were also described by

**TABLE 3 |** 2-locus linkage disequilibrium.

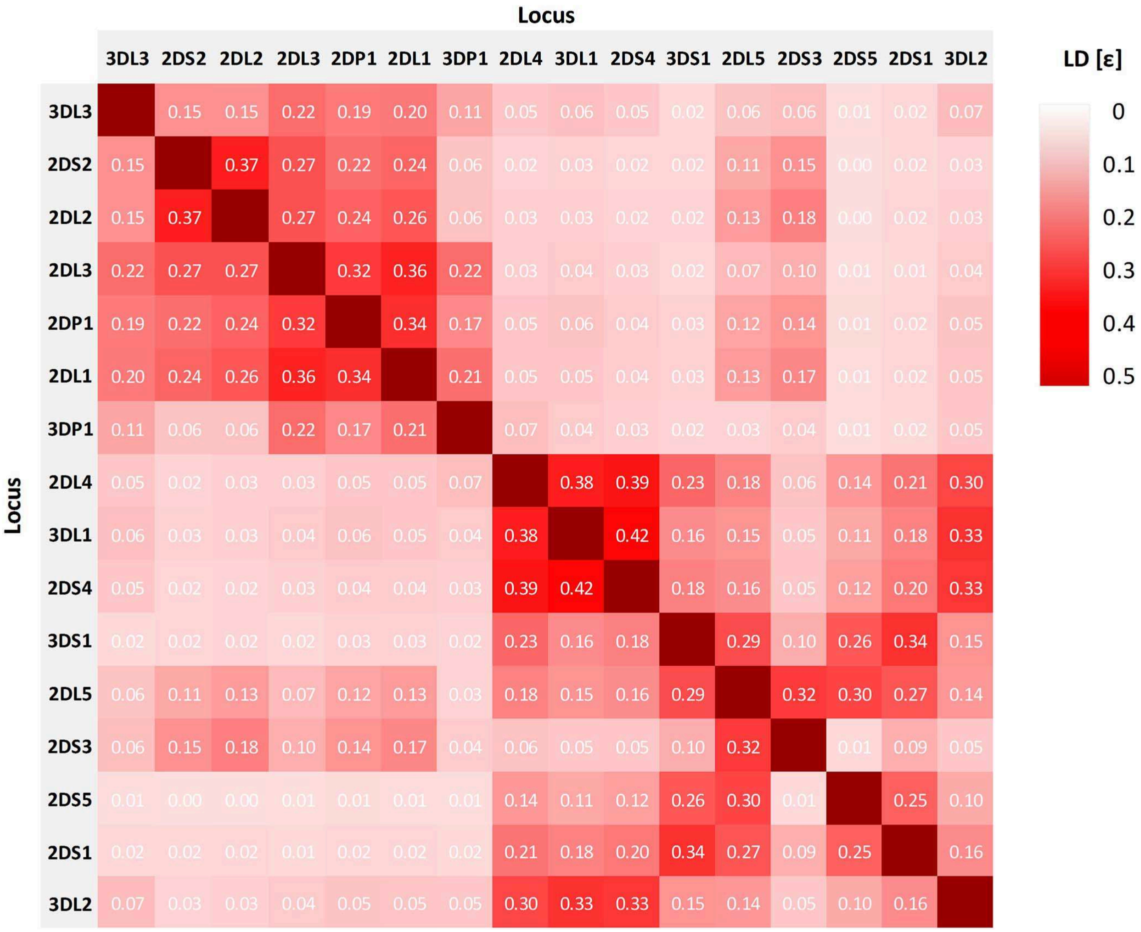
| Loci | Haplotype ab | f (ab) observed | f (a) | f (b) | D′ | f (ab) expected | *p* | $p^{(HB)}$ |
|---|---|---|---|---|---|---|---|---|
| 2DL1-3DP1 | 2DL1*001c-3DP1*006 | 0.1591 | 0.3080 | 0.1681 | 0.9225 | 0.0518 | 6.21E-13 | 2.33E-06 |
| 2DL2-2DP1 | 2DL2*001-2DP1*001 | 0.1038 | 0.1395 | 0.1095 | 0.9400 | 0.0153 | 4.78E-15 | 3.97E-06 |
| 2DL3-2DL1 | 2DL3*001-2DL1*003c | 0.3794 | 0.4107 | 0.4004 | 0.9109 | 0.1645 | 1.43E-22 | 3.30E-06 |
| 2DL3-2DL1 | 2DL3*002-2DL1*001c | 0.2916 | 0.3036 | 0.3080 | 0.9428 | 0.0935 | 1.05E-24 | 3.30E-06 |
| 2DL3-2DP1 | 2DL3*002-2DP1*003 | 0.2871 | 0.3036 | 0.3063 | 0.9214 | 0.0930 | 9.27E-24 | 3.35E-06 |
| 2DL3-3DP1 | 2DL3*002-3DP1*006 | 0.1661 | 0.3036 | 0.1681 | 0.9830 | 0.0510 | 3.36E-14 | 3.27E-06 |
| 2DL4-2DL5 | 2DL4*005-2DL5*001c | 0.1078 | 0.1882 | 0.1134 | 0.9388 | 0.0213 | 1.30E-13 | 1.58E-06 |
| 2DL4-2DS1 | 2DL4*005-2DS1*002c | 0.1056 | 0.1882 | 0.1127 | 0.9223 | 0.0212 | 4.05E-13 | 1.53E-06 |
| 2DL4-2DS4 | 2DL4*008-2DS4*006 | 0.1908 | 0.3646 | 0.1975 | 0.9466 | 0.0720 | 5.77E-13 | 1.54E-06 |
| 2DL4-2DS4 | 2DL4*001-2DS4*001c | 0.1899 | 0.2106 | 0.1971 | 0.9532 | 0.0415 | 9.42E-22 | 1.54E-06 |
| 2DL4-2DS4 | 2DL4*011-2DS4*010 | 0.1595 | 0.1772 | 0.1620 | 0.9810 | 0.0287 | 1.19E-20 | 1.54E-06 |
| 2DL4-2DS5 | 2DL4*005-2DS5*002 | 0.1272 | 0.1882 | 0.1354 | 0.9252 | 0.0255 | 2.58E-15 | 1.55E-06 |
| 2DL4-3DL1 | 2DL4*008-3DL1*004 | 0.1737 | 0.3646 | 0.1791 | 0.9523 | 0.0653 | 6.61E-12 | 1.66E-06 |
| 2DL4-3DL1 | 2DL4*011-3DL1*005 | 0.1699 | 0.1772 | 0.1724 | 0.9822 | 0.0306 | 6.13E-22 | 1.66E-06 |
| 2DL4-3DL1 | 2DL4*008-3DL1*001c | 0.1620 | 0.3646 | 0.1656 | 0.9656 | 0.0604 | 3.75E-11 | 1.66E-06 |
| 2DL4-3DL1 | 2DL4*001-3DL1*002 | 0.1039 | 0.2106 | 0.1104 | 0.9250 | 0.0233 | 6.27E-12 | 1.66E-06 |
| 2DL4-3DL2 | 2DL4*001-3DL2*002c | 0.1705 | 0.2106 | 0.1794 | 0.9367 | 0.0378 | 2.10E-19 | 1.43E-06 |
| 2DL4-3DL2 | 2DL4*005-3DL2*007 | 0.1658 | 0.1882 | 0.1767 | 0.9241 | 0.0333 | 7.40E-20 | 1.43E-06 |
| 2DL4-3DS1 | 2DL4*005-3DS1*013c | 0.1810 | 0.1882 | 0.1908 | 0.9524 | 0.0359 | 6.21E-22 | 1.64E-06 |
| 2DL5-2DS5 | 2DL5*001c-2DS5*002 | 0.1127 | 0.1134 | 0.1354 | 0.9925 | 0.0154 | 6.55E-17 | 1.21E-06 |
| 2DP1-2DL1 | 2DP1*002-2DL1*003c | 0.1967 | 0.2070 | 0.4004 | 0.9165 | 0.0829 | 1.88E-11 | 2.80E-06 |
| 2DP1-3DP1 | 2DP1*003-3DP1*006 | 0.1572 | 0.3063 | 0.1681 | 0.9064 | 0.0515 | 1.00E-12 | 2.77E-06 |
| 2DS2-2DL2 | 2DS2*001c-2DL2*003 | 0.1023 | 0.2180 | 0.1057 | 0.9592 | 0.0230 | 1.83E-11 | 4.65E-06 |
| 2DS4-3DL2 | 2DS4*001c-3DL2*002c | 0.1712 | 0.1971 | 0.1794 | 0.9429 | 0.0354 | 1.24E-20 | 1.10E-06 |
| 3DL1-2DS4 | 3DL1*004-2DS4*006 | 0.1766 | 0.1791 | 0.1975 | 0.9824 | 0.0354 | 1.15E-21 | 1.34E-06 |
| 3DL1-2DS4 | 3DL1*001c-2DS4*003 | 0.1634 | 0.1656 | 0.1810 | 0.9837 | 0.0300 | 3.67E-21 | 1.34E-06 |
| 3DL1-2DS4 | 3DL1*005-2DS4*010 | 0.1563 | 0.1724 | 0.1620 | 0.9574 | 0.0279 | 3.96E-20 | 1.34E-06 |
| 3DL1-2DS4 | 3DL1*002-2DS4*001c | 0.1060 | 0.1104 | 0.1971 | 0.9501 | 0.0218 | 1.24E-12 | 1.34E-06 |
| 3DL1-3DL2 | 3DL1*002-3DL2*002c | 0.1074 | 0.1104 | 0.1794 | 0.9662 | 0.0198 | 7.04E-14 | 1.26E-06 |
| 3DL3-2DL3 | 3DL3*002-2DL3*001 | 0.1203 | 0.1207 | 0.4107 | 0.9948 | 0.0496 | 1.54E-07 | 5.14E-05 |
| 3DL3-2DL3 | 3DL3*009-2DL3*001 | 0.1047 | 0.1082 | 0.4107 | 0.9455 | 0.0444 | 2.38E-06 | 5.17E-05 |
| 3DS1-2DS1 | 3DS1*013c-2DS1*002c | 0.1039 | 0.1908 | 0.1127 | 0.9033 | 0.0215 | 7.11E-13 | 1.24E-06 |

*Shown are the 32 significant cases which fulfill the filtering criteria (D′ ≥ 0.9, f (ab) ≥ 0.1, both loci present). f (ab) = haplotype frequency; f (a) and f (b), frequency of allele (groups) a and b, respectively; D′, relative linkage disequilibrium; p, p-value (Fisher's exact test); $p^{(HB)}$, p-value after Holm-Bonferroni correction.*

Hou et al. They could be assigned to four of the five postulated centromeric consensus structures. In the case of the telomeric part, 18 of our Top 20 allele group haplotypes were described in the previous study. They account for 7 out of 8 suggested telomeric haplotype structures.

The comparison of our results to *KIR* haplotype data of non-European individuals shows clearly less similarities. A recent study on *HLA* and *KIR* diversity in individuals from African populations reveals differences in the *KIR* HFs between the seven populations from Western, Central, and Eastern Africa (66). The African haplotypes furthermore show a different allelic distribution than those of our German cohort. Some of the alleles show exclusive occurrence in one of the geographically distinct cohorts. Alleles *3DL3*001* or *3DL1*002*, which are, for example, frequent in our sample, were not observed in the African samples. Vice versa, the African cohorts contain frequent alleles that are not found in our German haplotypes, e.g., *3DL3*005* or *3DL1*017*. The centromeric part of only seven and the telomeric part of 10 of our Top 20 allele group haplotypes are present in the African cohorts.

Two levels of linkage are visible in our LD analyses. On the one hand, observed LD in pairs of KIR loci clearly reflect the structural constraints of *KIR* haplotypes by echoing proximity of loci as given in the different CNPS haplotypes. On the other hand, the linkage of certain allele groups within these structural constraints indicates a non-stochastic distribution of alleles to particular haplotypes. The low deviation from linkage equilibrium between genes of the centromeric and the telomeric gene stretch supports the assumption of a recombination hot spot between *KIR3DP1* and *KIR2DL4*. Our findings on LD confirm data from previous research (44). Our set of 32 allele group LD pairs includes 20 of the 26 two-allele haplotypes that showed significant LD in the former study.



| | 3DL3 | 2DS2 | 2DL2 | 2DL3 | 2DP1 | 2DL1 | 3DP1 | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5 | 2DS3 | 2DS5 | 2DS1 | 3DL2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3DL3 | | 0.15 | 0.15 | 0.22 | 0.19 | 0.20 | 0.11 | 0.05 | 0.06 | 0.05 | 0.02 | 0.06 | 0.06 | 0.01 | 0.02 | 0.07 |
| 2DS2 | 0.15 | | 0.37 | 0.27 | 0.22 | 0.24 | 0.06 | 0.02 | 0.03 | 0.02 | 0.02 | 0.11 | 0.15 | 0.00 | 0.02 | 0.03 |
| 2DL2 | 0.15 | 0.37 | | 0.27 | 0.24 | 0.26 | 0.06 | 0.03 | 0.03 | 0.02 | 0.02 | 0.13 | 0.18 | 0.00 | 0.02 | 0.03 |
| 2DL3 | 0.22 | 0.27 | 0.27 | | 0.32 | 0.36 | 0.22 | 0.03 | 0.04 | 0.03 | 0.02 | 0.07 | 0.10 | 0.01 | 0.01 | 0.04 |
| 2DP1 | 0.19 | 0.22 | 0.24 | 0.32 | | 0.34 | 0.17 | 0.05 | 0.06 | 0.04 | 0.03 | 0.12 | 0.14 | 0.01 | 0.02 | 0.05 |
| 2DL1 | 0.20 | 0.24 | 0.26 | 0.36 | 0.34 | | 0.21 | 0.05 | 0.05 | 0.04 | 0.03 | 0.13 | 0.17 | 0.01 | 0.02 | 0.05 |
| 3DP1 | 0.11 | 0.06 | 0.06 | 0.22 | 0.17 | 0.21 | | 0.07 | 0.04 | 0.03 | 0.02 | 0.03 | 0.04 | 0.01 | 0.02 | 0.05 |
| 2DL4 | 0.05 | 0.02 | 0.03 | 0.03 | 0.05 | 0.05 | 0.07 | | 0.38 | 0.39 | 0.23 | 0.18 | 0.06 | 0.14 | 0.21 | 0.30 |
| 3DL1 | 0.06 | 0.03 | 0.03 | 0.04 | 0.06 | 0.05 | 0.04 | 0.38 | | 0.42 | 0.16 | 0.15 | 0.05 | 0.11 | 0.18 | 0.33 |
| 2DS4 | 0.05 | 0.02 | 0.02 | 0.03 | 0.04 | 0.04 | 0.03 | 0.39 | 0.42 | | 0.18 | 0.16 | 0.05 | 0.12 | 0.20 | 0.33 |
| 3DS1 | 0.02 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.02 | 0.23 | 0.16 | 0.18 | | 0.29 | 0.10 | 0.26 | 0.34 | 0.15 |
| 2DL5 | 0.06 | 0.11 | 0.13 | 0.07 | 0.12 | 0.13 | 0.03 | 0.18 | 0.15 | 0.16 | 0.29 | | 0.32 | 0.30 | 0.27 | 0.14 |
| 2DS3 | 0.06 | 0.15 | 0.18 | 0.10 | 0.14 | 0.17 | 0.04 | 0.06 | 0.05 | 0.05 | 0.10 | 0.32 | | 0.01 | 0.09 | 0.05 |
| 2DS5 | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.14 | 0.11 | 0.12 | 0.26 | 0.30 | 0.01 | | 0.25 | 0.10 |
| 2DS1 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.21 | 0.18 | 0.20 | 0.34 | 0.27 | 0.09 | 0.25 | | 0.16 |
| 3DL2 | 0.07 | 0.03 | 0.03 | 0.04 | 0.05 | 0.05 | 0.05 | 0.30 | 0.33 | 0.33 | 0.15 | 0.14 | 0.05 | 0.10 | 0.16 | |

**FIGURE 4 |** Overall linkage disequilibrium of 16 *KIR* loci in genomic orientation as designated by the normalized entropy difference ε. The value of ε ranges between 0 and 1, with larger values indicating stronger LD. Values of ε are displayed in the tiles.

The genetic linkage of alleles within given haplotype structures reflected in the LD analysis is confirmed when allele group haplotypes are clustered by their underlying copy number haplotype. The analysis reveals two interesting results: First, allelic variability in our cohort is found to be higher in A than in B haplotypes. This observation agrees with previous descriptions (48) and is also in accord with findings of lower overall allelic diversity in the activating compared to the inhibitory loci (44, 53). And second, we observe conserved allelic correlations specific to five distinct centromeric and telomeric haplotype motifs (cA01, cB01, cB02, tA01, and tB01). The detected differences in allelic variability and the correlation of certain alleles to A and B haplotype motifs also comprises the framework genes. These results indicate a *KIR* gene inheritance in closely linked blocks with a recombination hot spot between the genes *KIR3DP1* and *KIR2DL4*. The very particular allelic linkage within the *KIR2DL5*~*KIR2DS3/2DS5* haplotype enabled us to assign this gene cluster to centromeric or telomeric position in our KIR haplotypes even though our typing method did not allow a differentiation between the two loci KIR2DL5A and KIR2DL5B. Beyond that, the observed allele group haplotype patterns of the telomeric and centromeric *KIR2DL5*~*KIR2DS3/2DS5* clusters correspond to those described in previous publications on individuals of European ancestry (52, 60).

The knowledge of family relations in our cohort substantially reduced the number of possible diplotypes per individual. This had two consequences. First, it increased the number of individuals who were excluded from the *KIR* HF estimation due to the absence of a valid combination of CNPS haplotypes that explained their diplotypes. This demonstrates, on the one hand, the incompleteness of our used CNPS haplotype list. On the other hand, it reveals the limitation of *KIR* HF estimation without family context, where, in consequence, the same *KIR* genotypes were described with at least one incorrectly assigned pair of CNPS haplotypes. And second, the reduced number of possible diplotypes per individual provided a reasonable basis for a successful execution of an EM algorithm. Apart from reducing phasing ambiguities, the family context further offers the potential to detect typing inaccuracies and recombination events that would remain hidden otherwise.

Given the multiple constraints we applied to our *KIR* data in order to estimate allele group *KIR* HFs, several sources of possible bias to our results have to be considered. One important bias may be caused by the restriction of *KIR* haplotypes to a limited copy number pattern set. For 10.6% of the parents in our cohort, the deduced diplotypes could not be explained by a pair of haplotypes from the CNPS, indicating further *KIR* complexity on the copy number level and the need for extension of the CNPS. The estimation of partial (centromeric and telomeric) haplotypes of the same cohort led to a reduction of this percentage of excluded individuals. This indicates that the missing structural diversification of the pattern haplotypes is mostly limited to variation in sub-ranges of the KIR gene complex. Thorough analysis of the diplotypes of the excluded individuals by alternative typing methods beyond the routine high-throughput *KIR* genotyping in the context of unrelated HSCT donor registration (53) will be of great importance for the future refinement of our approach. Such extended analyses could also detect hybrid genes that are not described in the IPD-*KIR* database to which our routine typing method is blind, but go beyond the scope of the present study. Further bias may be introduced by sequencing or sequence interpretation errors. However, the impact of this potential bias should be minimal with our approach because the low overall error rate of the NGS high throughput platform was further reduced by verification of the *KIR* data within the families. The accuracy of the high throughput *KIR* analysis in a curated sample was found to exceed 99% (53). In the 13 remaining cases of mismatching loci and one presumable case of recombination in our cohort, where typing errors could not be excluded as a cause without re-typing of the respective family members, parents were excluded from the HF estimation. Moreover, the restriction of the cohort to individuals with German origin needs to be treated with some caution. This information was collected via self-assessment of the volunteer stem cell donors during the registration process. The perception of the term "origin" is quite individual (67) and in many cases it is simply difficult to describe an individual's origin by means of one single country code. Finally, family lineage was not known for the individuals of our cohort but deduced from concordance of surname and address and a defined age difference between presumed parents and offspring. However, the verification of genetic relationship via the *HLA* genes and exclusion of families with any mismatching loci in the *KIR* genes from the final HF estimation should compensate for lack of direct family information.

In conclusion, our approach yielded a set of 551 allele group *KIR* haplotype frequencies from a German cohort that is in good accordance with results from other groups. We provide additional data on the diversification of *KIR* haplotypes by inclusion of allelic polymorphisms of pseudogenes *KIR2DP1* and *KIR3DP1*. The use of family information during diplotype deduction allowed the exclusion of incorrect phasing variants. Our *KIR* haplotype frequencies reveal relations between *KIR* copy number haplotypes and allele frequencies, which will be a valuable basis for future research. The application of this approach, e.g., to larger cohorts of different ethnic origin, will further broaden our knowledge and understanding of the very complex nature of the *KIR* genes.

## RESEARCH DATA

DNA was extracted from blood samples or buccal swabs with the informed consent of the donors. Research data used in this publication was collected from the data subjects and processed on the basis of an informed consent in accordance with the EU Data Protection Regulation (EU-GDPR). Data subjects agreed to their data being processed for scientific studies, in particular with the aim to improving the treatment of patients with blood cancer and other life threatening diseases. The publication itself does not include identifiable personal data.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## AUTHOR CONTRIBUTIONS

JSa and US conceived the project. US designed and implemented algorithms, carried out analyses, and wrote the first draft of the manuscript. GS, CM, AH, MK, DS, and JP analyzed and interpreted KIR genotyping data. All authors contributed to manuscript revision, read, and approved the submitted version.

## ACKNOWLEDGMENTS

We thank all the altruistic volunteers who register as potential donors to help strangers in need of a stem cell transplant. We would also like to thank Beatrix Willburger for valuable advice during the implementation of the algorithm. We thank the reviewers for their valuable advice and comments.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00429/full#supplementary-material

## REFERENCES

1. Marsh SGE, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Hum Immunol*. (2003) 64:648–54. doi: 10.1016/S0198-8859(03)00067-3
2. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci USA*. (2000) 97:4778–83. doi: 10.1073/pnas.080588597
3. Middleton D, Gonzelez F. The extensive polymorphism of KIR genes. *Immunology*. (2010) 129:8–19. doi: 10.1111/j.1365-2567.2009.03208.x
4. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes

of the human KIR genes encoding NK cell receptors. *Genome Res*. (2012) 22:1845–54. doi: 10.1101/gr.137976.112

5. Robinson J, Waller MJ, Stoehr P, Marsh SG. IPD–the immuno polymorphism database. *Nucleic Acids Res*. (2005) 33:D523–6. doi: 10.1093/nar/gki032
6. Augusto DG, Norman PJ, Dandekar R, Hollenbach JA. Fluctuating and geographically specific selection characterize rapid evolution of the human KIR region. *Front Immunol*. (2019) 10:989. doi: 10.3389/fimmu.2019.00989
7. Shilling HG, Lienert-Weidenbach K, Valiante NM, Uhrberg M, Parham P. Evidence for recombination as a mechanism for KIR diversification. *Immunogenetics*. (1998) 48:413–6. doi: 10.1007/s002510050453
8. Martin MP, Bashirova A, Traherne J, Trowsdale J, Carrington M. Cutting edge: expansion of the KIR locus by unequal crossing over. *J Immunol*. (2003) 171:2192–5. doi: 10.4049/jimmunol.171.5.2192
9. Abi-Rached L, Parham P. Natural selection drives recurrent formation of activating killer cell immunoglobulin-like receptor and Ly49 from inhibitory homologues. *J Exp Med*. (2005) 201:1319–32. doi: 10.1084/jem.20042558
10. Gomez-Lozano N, Estefania E, Williams F, Halfpenny I, Middleton D, Solis R, et al. The silent KIR3DP1 gene (CD158c) is transcribed and might encode a secreted receptor in a minority of humans, in whom the KIR3DP1, KIR2DL4 and KIR3DL1/KIR3DS1 genes are duplicated. *Eur J Immunol*. (2005) 35:16–24. doi: 10.1002/eji.200425493
11. Norman PJ, Abi-Rached L, Gendzekhadze K, Hammond JA, Moesta AK, Sharma D, et al. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. *Genome Res*. (2009) 19:757–69. doi: 10.1101/gr.085738.108
12. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet*. (2010) 19:737–51. doi: 10.1093/hmg/ddp538
13. Ordonez D, Gomez-Lozano N, Rosales L, Vilches C. Molecular characterisation of KIR2DS2*005, a fusion gene associated with a shortened KIR haplotype. *Genes Immun*. (2011) 12:544–51. doi: 10.1038/gene.2011.35
14. Roe D, Vierra-Green C, Pyo CW, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun*. (2017). 18:127–34. doi: 10.1038/gene.2017.10
15. Bruijnesteijn J, van der Wiel MKH, de Groot N, Otting N, de Vos-Rouweler AJM, Lardy NM, et al. Extensive alternative splicing of KIR transcripts. *Front Immunol*. (2018) 9:2846. doi: 10.3389/fimmu.2018.02846
16. Vivier E, Ugolini S, Blaise D, Chabannon C, Brossay L. Targeting natural killer cells and natural killer T cells in cancer. *Nat Rev Immunol*. (2012) 12:239–52. doi: 10.1038/nri3174
17. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol*. (2013) 13:133–44. doi: 10.1038/nri3370
18. Cooley S, Parham P, Miller JS. Strategies to activate NK cells to prevent relapse and induce remission following hematopoietic stem cell transplantation. *Blood*. (2018) 131:1053–62. doi: 10.1182/blood-2017-08-752170
19. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, et al. Human diversity in killer cell inhibitory receptor genes. *Immunity*. (1997) 7:753–63. doi: 10.1016/S1074-7613(00)80394-5
20. Faure M, Long EO. KIR2DL4 (CD158d), an NK cell-activating receptor with inhibitory potential. *J Immunol*. (2002) 168:6208–14. doi: 10.4049/jimmunol.168.12.6208
21. Rajagopalan S, Long EO. KIR2DL4 (CD158d): An activation receptor for HLA-G. *Front Immunol*. (2012) 3:258. doi: 10.3389/fimmu.2012.00258
22. Gómes-Lozano N, de Pablo R, Puente S, Vilches C. Recognition of HLA-G by the NK cell receptor KIR2DL4 is not essential for human reproduction. *Eur J Immunol*. (2003) 33:639–44. doi: 10.1002/eji.200323741
23. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood*. (2007) 110:4576–83. doi: 10.1182/blood-2007-06-097386
24. Pidala J, Lee SJ, Ahn KW, Spellman S, Wang HL, Aljurf M, et al. Nonpermissive HLA-DPB1 mismatch increases mortality after myeloablative unrelated allogeneic hematopoietic cell transplantation. *Blood*. (2014) 124:2596–606. doi: 10.1182/blood-2014-05-576041
25. Dehn J, Spellman S, Hurley CK, Shaw BE, Barker JN, Burns LJ, et al. Selection of unrelated donors and cord blood units for hematopoietic cell transplantation: guidelines from NMDP/CIBMTR. *Blood*. (2019) 134:924–34. doi: 10.1182/blood.2019001212
26. Ullah MA, Hill GR, Tey SK. Functional Reconstitution of natural killer cells in allogeneic hematopoietic stem cell transplantation. *Front Immunol*. (2016) 7:144. doi: 10.3389/fimmu.2016.00144
27. Ruggeri L, Capanni M, Casucci M, Volpi I, Tosti A, Perruccio K, et al. Role of natural killer cell alloreactivity in HLA-mismatched hematopoietic stem cell transplantation. *Blood*. (1999) 94:333–9. doi: 10.1182/blood.V94.1.333.413a31_333_339
28. Ruggeri L, Capanni M, Urbani E, Perruccio K, Shlomchik WD, Tosti A, et al. Effectiveness of donor natural killer cell alloreactivity in mismatched hematopoietic transplants. *Science*. (2002) 295:2097–100. doi: 10.1126/science.1068440
29. Giebel S, Locatelli F, Lamparelli T, Velardi A, Davies S, Frumento G, et al. Survival advantage with KIR ligand incompatibility in hematopoietic stem cell transplantation from unrelated donors. *Blood*. (2003) 102:814–9. doi: 10.1182/blood-2003-01-0091
30. Giebel S, Locatelli F, Wojnar J, Velardi A, Mina T, Giorgiani G, et al. Homozygosity for human leucocyte antigen-C ligands of KIR2DL1 is associated with increased risk of relapse after human leucocyte antigen-C-matched unrelated donor haematopoietic stem cell transplantation. *Br J Haematol*. (2005) 131:483–6. doi: 10.1111/j.1365-2141.2005.05797.x
31. Miller JS, Cooley S, Parham P, Farag SS, Verneris MR, McQueen KL, et al. Missing KIR ligands are associated with less relapse and increased graft-versus-host disease (GVHD) following unrelated donor allogeneic HCT. *Blood*. (2007) 109:5058–61. doi: 10.1182/blood-2007-01-065383
32. Heidenreich S, Kroger N. Reduction of relapse after unrelated donor stem cell transplantation by KIR-based graft selection. *Front Immunol*. (2017) 8:41. doi: 10.3389/fimmu.2017.00041
33. Cooley S, Trachtenberg E, Bergemann TL, Saeteurn K, Klein J, Le CT, et al. Donors with group B KIR haplotypes improve relapse-free survival after unrelated hematopoietic cell transplantation for acute myelogenous leukemia. *Blood*. (2009) 113:726–32. doi: 10.1182/blood-2008-07-171926
34. Venstrom JM, Pittari G, Gooley TA, Chewning JH, Spellman S, Haagenson M, et al. HLA-C-dependent prevention of leukemia relapse by donor activating KIR2DS1. *N Engl J Med*. (2012) 367:805–16. doi: 10.1056/NEJMoa1200503
35. Boudreau JE, Giglio F, Gooley TA, Stevenson PA, Le Luduec J-B, Shaffer BC, et al. KIR3DL1/HLA-B subtypes govern acute myelogenous leukemia relapse after hematopoietic cell transplantation. *J Clin Oncol*. (2017). doi: 10.1200/JCO.2016.70.7059
36. Schetelig J, Baldauf H, Koster L, Kuxhausen M, Heidenreich F, De Wreede LC, et al. Does donor KIR-genotype impact outcome after unrelated hematopoietic stem cell transplantation for myelodysplastic syndromes or secondary acute myeloid leukemia? *Bone Marrow Transplant*. (2019) 54:561–2. doi: 10.1038/s41409-019-0559-4
37. Schetelig J, Baldauf H, Heidenreich F, Massalski C, Frank S, Sauter J, et al. External validation of models for KIR2DS1/KIR3DL1-informed Selection of hematopoietic cell donors fails. *Blood*. (2020). doi: 10.1182/blood.2019002887. [Epub ahead of print].
38. Urban C, Schmidt AH, Hofmann JA. (2020). Hap-E Search 2.0: improving the performance of a probabilistic donor-recipient matching algorithm based on haplotype frequencies. *Front Med*. 7:32. doi: 10.3389/fmed.2020.00032
39. Bochtler W, Gragert L, Patel ZI, Robinson J, Steiner D, Hofmann JA, et al. A comparative reference study for the validation of *HLA*-matching algorithms in the search for allogeneic hematopoietic stem cell donors and cord blood units. *HLA*. (2016) 87:439–48. doi: 10.1111/tan.12817
40. Schmidt AH, Sauter J, Pingel J, Ehninger G. Toward an optimal global stem cell donor recruitment strategy. *PLoS ONE*. (2014) 9:e86605. doi: 10.1371/journal.pone.0086605
41. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med*. (2007) 4:e8. doi: 10.1371/journal.pmed.0040008
42. Joris MM, Lankester AC, von dem Borne PA, Kuball J, Bierings M, Cornelissen JJ, et al. The impact of frequent HLA haplotypes in high linkage disequilibrium on donor search and clinical outcome after unrelated haematopoietic SCT. *Bone Marrow Transplant*. (2013) 48:483–90. doi: 10.1038/bmt.2012.189
43. Buhler S, Baldomero H, Ferrari-Lacraz S, Nunes JM, Sanchez-Mazas A, Massouridi-Levrat S, et al. High-resolution HLA phased haplotype

frequencies to predict the success of unrelated donor searches and clinical outcome following hematopoietic stem cell transplantation. *Bone Marrow Transplant*. (2019) 54:1701–9. doi: 10.1038/s41409-019-0520-6

44. Vierra-Green C, Roe D, Hou L, Hurley CK, Rajalingam R, Reed E, et al. Allele-level haplotype frequencies and pairwise linkage disequilibrium for 14 KIR loci in 506 European-American individuals. *PLoS ONE.* (2012) 7:e47491. doi: 10.1371/journal.pone.0047491
45. Vierra-Green C, Roe D, Jayaraman J, Trowsdale J, Traherne J, Kuang R, et al. Estimating KIR haplotype frequencies on a cohort of 10,000 individuals: a comprehensive study on population variations, typing resolutions, and reference haplotypes. *PLoS ONE.* (2016) 11:e0163973. doi: 10.1371/journal.pone.0163973
46. Hsu KC, Liu XR, Selvakumar A, Mickelson E, O'Reilly RJ, Dupont B. Killer Ig-Like receptor haplotype analysis by gene content: evidence for genomic diversity with a minimum of six basic framework haplotypes, each with multiple subsets. *J Immunol*. (2002) 169:5118–29. doi: 10.4049/jimmunol.169.9.5118
47. Uhrberg M, Parham P, Wernet P. Definition of gene content for nine common group B haplotypes of the caucasoid population: KIR haplotypes contain between seven and eleven KIR genes. *Immunogenetics*. (2002) 54:221–9. doi: 10.1007/s00251-002-0463-7
48. Pyo CW, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different patterns of evolution in the centromeric and telomeric regions of group A and B haplotypes of the human killer cell Ig-like receptor locus. *PLoS ONE.* (2010) 5:e15115. doi: 10.1371/journal.pone.0015115
49. Middleton D, Meenagh A, Gourraud PA. KIR haplotype content at the allele level in 77 Northern Irish families. *Immunogenetics*. (2007) 59:145–58. doi: 10.1007/s00251-006-0181-7
50. Pyo CW, Wang R, Vu Q, Cereb N, Yang SY, Duh FM, et al. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics*. (2013) 14:89. doi: 10.1186/1471-2164-14-89
51. Shilling HG, Guethlein LA, Cheng NW, Gardiner CM, Rodriguez R, Tyan D, et al. Allelic polymorphism synergizes with variable gene content to individualize human KIR genotype. *J Immunol*. (2002) 168:2307–15. doi: 10.4049/jimmunol.168.5.2307
52. Hou L, Chen M, Ng J, Hurley CK. Conserved KIR allele-level haplotypes are altered by microvariation in individuals with European ancestry. *Genes Immun*. (2012) 13:47–58. doi: 10.1038/gene.2011.52
53. Wagner I, Schefzyk D, Pruschke J, Schöfl G, Schöne B, Gruber N, et al. Allele-level KIR genotyping of more than a million samples: workflow, algorithm, and observations. *Front Immunol*. (2018) 9:2843. doi: 10.3389/fimmu.2018.02843
54. Schmidt AH, Lange V, Hofmann JA, Schetelig J, Pingel J. KIR genotyping data of more than 3 million individuals are available for global unrelated stem cell donor searches. (comment to: Weisdorf D, Cooley S, Wang T, et al. KIR donor selection: feasibility in identifying better donors). *Biol Blood Marrow Transplant*. (2019) 25:e39–40. doi: 10.1016/j.bbmt.2018.09.020
55. Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*. (2014) 15:63. doi: 10.1186/1471-2164-15-63
56. Schöfl G, Lang K, Quenzel P, Böhme I, Sauter J, Hofmann JA, et al. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC Genomics*. (2017) 18:161. doi: 10.1186/s12864-017-3575-z
57. Lang K, Wagner I, Schöne B, Schöfl G, Birkner K, Hofmann JA, et al. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genomics*. (2016) 17:374. doi: 10.1186/s12864-016-2687-1
58. Solloch UV, Lang K, Lange V, Böhme I, Schmidt AH, Sauter J. Frequencies of gene variant CCR5-Δ32 in 87 countries based on next-generation sequencing of 1.3 million individuals sampled from 3 national DKMS donor centers. *Hum Immunol.* (2017) 78:710–7. doi: 10.1016/j.humimm.2017.10.001
59. Klussmeier A, Massalski C, Putke K, Schäfer G, Sauter J, Schefzyk D, et al. High-throughput MICA/B genotyping of over two million samples: workflow and allele frequencies. *Front Immunol.* (2020) 11:314. doi: 10.3389/fimmu.2020.00314
60. Ordonez D, Meenagh A, Gomez-Lozano N, Castano J, Middleton D, Vilches C. Duplication, mutation and recombination of the human orphan gene KIR2DS3 contribute to the diversity of KIR haplotypes. *Genes Immun*. (2008) 9:431–7. doi: 10.1038/gene.2008.34
61. Cisneros E, Moraru M, Gómez-Lozano N, López-Botet M, Vilches C. KIR2DL5: an orphan inhibitory receptor displaying complex patterns of polymorphism and expression. *Front Immunol.* (2012) 3:289. doi: 10.3389/fimmu.2012.00289
62. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*. (1995) 12:921–7.
63. Lewontin RC. The interaction of selection and linkage. I general considerations; heterotic models. *Genetics*. (1964) 49:49–67.
64. Nothnagel M, Fürst R, Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered*. (2002) 54:186–98. doi: 10.1159/000070664
65. Okada Y. eLD: Entropy-based linkage disequilibrium index between multiallelic sites. *Hum Genome Var*. (2018) 5:29. doi: 10.1038/s41439-018-0030-x
66. Nemat-Gorgani N, Guethlein LA, Henn BM, Norberg SJ, Chiaroni J, Sikora M, et al. Diversity of KIR, HLA class I, and their interactions in seven populations of Sub-Saharan Africans. *J Immunol*. (2019) 202:2636–47. doi: 10.4049/jimmunol.1801586
67. Hollenbach JA, Saperstein A, Albrecht M, Vierra-Green C, Parham P, Norman PJ, et al. Race, ethnicity and ancestry in unrelated transplant matching for the national marrow donor program: a comparison of multiple forms of self-identification with genetics. *PLoS ONE.* (2015) 10:e0135960. doi: 10.1371/journal.pone.0135960

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 17 March 2020
doi: 10.3389/fimmu.2020.00440

# Haplotype-Based Analysis of *KIR*-Gene Profiles in a South European Population—Distribution of Standard and Variant Haplotypes, and Identification of Novel Recombinant Structures

*Elisa Cisneros [1], Manuela Moraru [1], Natalia Gómez-Lozano [1], Aura Muntasell [2], Miguel López-Botet [2,3] and Carlos Vilches [1]**

[1] Immunogenetics and Histocompatibility, Instituto de Investigación Sanitaria Puerta de Hierro Segovia de Arana, Madrid, Spain, [2] Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain, [3] Department of Experimental and Health Sciences, University Pompeu Fabra, Barcelona, Spain

OPEN ACCESS

**Edited by:**
Martin Maiers,
National Marrow Donor Program,
United States

**Reviewed by:**
Karl-Johan Malmberg,
Oslo University Hospital, Norway
Francesco Colucci,
University of Cambridge,
United Kingdom

***Correspondence:**
Carlos Vilches
carlos.vilches@yahoo.com

**Specialty section:**
This article was submitted to
NK and Innate Lymphoid Cell Biology,
a section of the journal
Frontiers in Immunology

**Received:** 19 January 2020
**Accepted:** 26 February 2020
**Published:** 17 March 2020

**Citation:**
Cisneros E, Moraru M,
Gómez-Lozano N, Muntasell A,
López-Botet M and Vilches C (2020)
Haplotype-Based Analysis of
KIR-Gene Profiles in a South
European Population—Distribution of
Standard and Variant Haplotypes, and
Identification of Novel Recombinant
Structures. Front. Immunol. 11:440.
doi: 10.3389/fimmu.2020.00440

Inhibitory Killer-cell Immunoglobulin-like Receptors (KIR) specific for HLA class I molecules enable human natural killer cells to monitor altered antigen presentation in pathogen-infected and tumor cells. *KIR* genes display extensive copy-number variation and allelic polymorphism. They organize in a series of variable arrangements, designated *KIR* haplotypes, which derive from duplications of ancestral genes and sequence diversification through point mutation and unequal crossing-over events. Genomic studies have established the organization of multiple *KIR* haplotypes—many of them are fixed in most human populations, whereas variants of those have less certain distributions. Whilst *KIR*-gene diversity of many populations and ethnicities has been explored superficially (frequencies of individual genes and presence/absence profiles), less abundant are in-depth analyses of how such diversity emerges from *KIR*-haplotype structures. We characterize here the genetic diversity of KIR in a sample of 414 Spanish individuals. Using a parsimonious approach, we manage to explain all 38 observed *KIR*-gene profiles by homo- or heterozygous combinations of six fixed centromeric and telomeric motifs; of six variant gene arrangements characterized previously by us and others; and of two novel haplotypes never detected before in Caucasoids. Associated to the latter haplotypes, we also identified the novel transcribed *KIR2DL5B*0020202* allele, and a chimeric *KIR2DS2/KIR2DL3* gene (designated *KIR2DL3*033*) that challenges current criteria for classification and nomenclature of *KIR* genes and haplotypes.

**Keywords: copy-number variation, genes, haplotypes, KIR, NK cells, polymorphism**

## INTRODUCTION

Human killer-cell immunoglobulin-like receptors (KIR) are a diverse and polymorphic family of glycoproteins that convey inhibitory or activating signals to subpopulations of NK and T lymphocytes upon recognition of their ligands, mainly HLA class I allotypes (1). In coordination with multiple other activating and inhibitory receptors for HLA class I and non-HLA molecules, KIR regulate the function of cytotoxic lymphocytes, providing them with a capacity to sense modifications of HLA expression on potential target cells (2, 3).

KIR repertoires expressed by NK cells of different individuals display a conspicuous phenotypic and functional diversity, genetically determined in its greatest part. KIR are encoded in a ∼100–250 Kbp complex on chromosome 19q13.4, where variable combinations of 15 *KIR* genes and 2 pseudogenes arrange in a head-to-tail orientation, separated by intergenic regions of only ∼2.5 Kbp (4). Another ancestral gene, *KIR3DX1* (5), lays ∼178 Kbp upstream of the *KIR* complex, in the middle of the *LILR-LAIR* gene cluster (**Figure 1**).

The *KIR* complex is extremely diverse due to allelic polymorphism and gene copy-number variation (CNV). Only three "framework" regions of the *KIR* complex are relatively well-conserved in their gene content: the genes at the 5′ and 3′ ends (*KIR 3DL3* and *3DL2*, respectively), and a central cluster formed by *KIR 3DP1* and *2DL4*. These framework *KIR* genes define the limits of two intervals, centromeric (5′) and telomeric (3′), containing variable combinations of the other genes (4, 6–10). Certain gene arrangements or "motifs" are particularly common within each of those intervals (**Figure 1**); in turn, the different centromeric and telomeric gene motifs are seen in any combinatorial association, possibly owing to a recombination hot-spot between *KIR 3DP1* and *2DL4*.

Restriction-fragment length polymorphism studies published in 1997 sorted *KIR* genotypes into categories "A" and "B," based on variable presence of a 24 Kbp-long *Hind*III band, later shown to derive from the *KIR2DL5* gene (11, 12). This definition was then refined and adapted (13), so that "A haplotype" now officially designates a nearly fixed combination of seven genes and pseudogenes, encoding the HLA-C-specific KIR 2DL3 and 2DL1 in the variable centromeric interval, and 3DL1 and 2DS4 in the telomeric one. In contrast, "B" designates collectively a vast array of haplotypes bearing any additional *KIR* gene (even when they also have, as it often happens, parts of an A haplotype). Of immunologic relevance, the "A" haplotype encodes inhibitory KIR for all known HLA ligands; and, at most, a single activating KIR expressed on the NK-cell surface, KIR2DS4 being often represented by an aberrant allele (14). In contrast, "B" haplotypes are distinguished by one or more of the following features: they encode several activating KIR; lack one or more genes for the aforementioned inhibitory KIR; and/or carry *KIR2DL5*, of uncertain biological role (15). This diversity appears to influence many human health conditions (16).

Certain neighbor *KIR* genes tend to appear strongly linked in the same haplotype. Noteworthy among those is the pair formed by *KIR 2DL5* and either *2DS3* or *2DS5*, these being inherited like allotypes of the same locus (17). This *KIR 2DL5–2DS3/S5* cluster duplicated and diversified jointly during human evolution, and is seen on either or both of the centromeric or the telomeric intervals of many B haplotypes (8, 15, 17, 18). Existence of two such long and highly similar stretches of sequence favored further asymmetric recombination between the paralogous regions. This resulted in expanded and shortened haplotypes bearing tandem duplications or deletions of the intervening genes, as shown by us and others (9, 10, 15, 19–24). Additional consequences of asymmetric recombination are novel fusion genes encoding chimeric KIR that blend structural and functional features of their parent receptors.

Initial studies of human *KIR* genotypes faced this vast polymorphism without previous knowledge of the structure and forms of variation of the *KIR*-gene complex, revealing multiple *KIR*-gene and allele profiles in different individuals, which could give a false impression of randomness (11, 25). Order, in the form of knowledge on common and variant patterns of association between *KIR* genes and alleles, emerged from subsequent studies of population groups and families; and from phasing and physical mapping of partial and complete *KIR*-gene haplotypes by DNA sequencing (4, 6–10, 26–29). Those studies were complemented by others focused on estimation of CNV and allelic diversity (23, 24, 30–39). In parallel, *KIR*-gene profiles were studied in many healthy and diseased populations and ethnic groups worldwide (40). However, many those population studies have benefited surprisingly little from knowledge gained in the last years on the structure and patterns of variability of the *KIR* gene complex, and, for many human populations, only collections of *KIR*-gene profiles and superficial analyses of the basic variations are available. Here, we have applied current knowledge of *KIR*-gene arrangements to a comprehensive analysis of the gene profiles observed in a European Mediterranean population.

# MATERIALS AND METHODS

## Samples

Genomic DNA was isolated using standard methods from peripheral blood or mononuclear cell (PBMC) suspensions, obtained by Ficoll-Hypaque density gradient centrifugation (Lymphoprep, Axis-Shield PoC AS, Oslo, Norway), from 414 unrelated voluntary donors recruited in our centers in Madrid and Barcelona, mostly of Caucasoid origin; only known exceptions were two donors of mixed Hispanic/Amerindian ancestry, neither of whom contributed novel or variant gene profiles. Complementary DNA was synthesized with the AffinityScript Multiple Temperature cDNA Synthesis Kit (Agilent Technologies, Santa Clara, CA, USA) from 400 ng of total RNA, extracted from PBMCs of selected donors using the RNeasy Plus Mini kit (Qiagen GmbH, D-40724, Hilden, Germany).

## *KIR* Genotyping

*KIR* genes, structural variants of *KIR 2DS4*, *2DL5,* and *3DP1*, and the hybrid alleles *2DS2**005* and *3DP1**004* were typed by PCR with sequence-specific primers (SSP), as previously described (22, 41–43). Non-standard *KIR*-gene profiles were confirmed utilizing a commercial reverse oligonucleotide probe-hybridization method based on the Luminex xMap Technology (LabType SSO Test, One Lambda Inc., Canoga Park, CA). To verify presence of an expanded haplotype in a *3DP1**004*$^{-ve}$ donor, existence of three *KIR3DL1/S1* alleles was verified by sequence-based typing of exons 3–5 in two overlapping amplicons. Each of these was generated with Advantage-2 polymerase (BD-Clontech, Palo Alto, CA, USA)

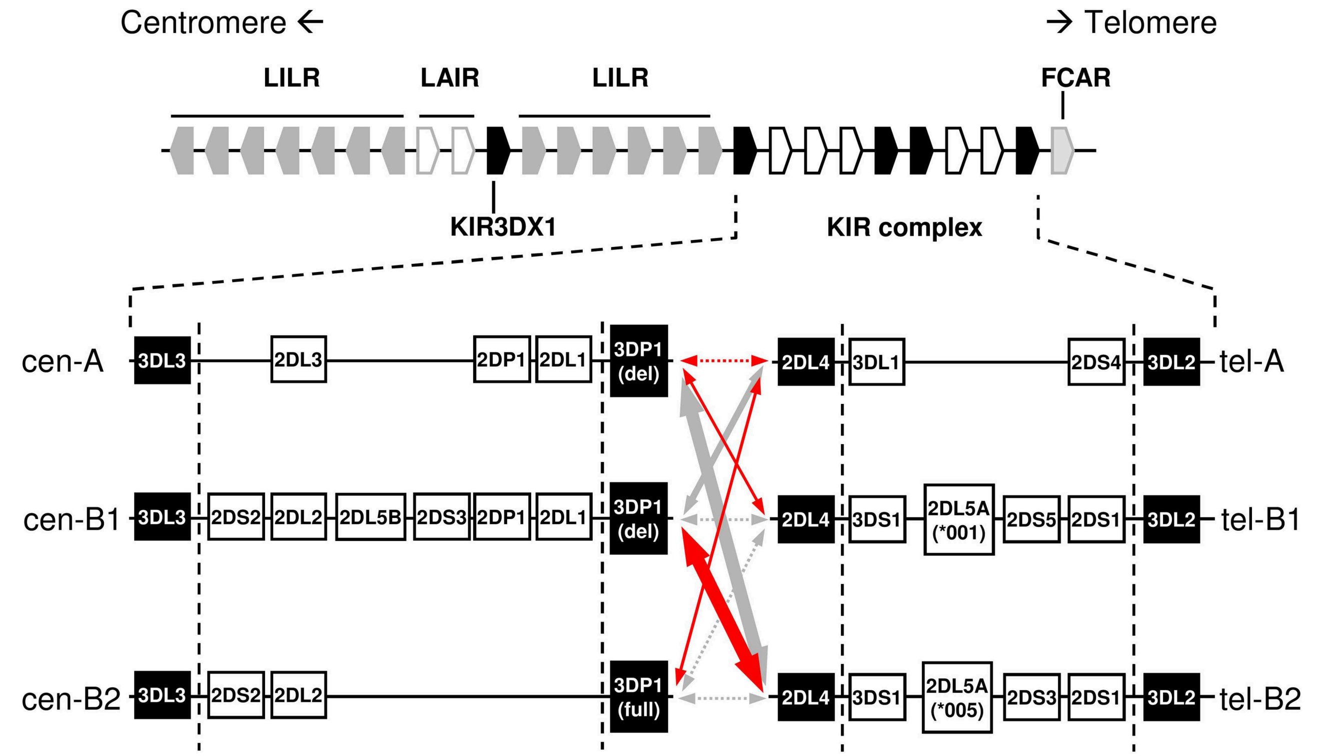


**FIGURE 1 |** Centromeric and telomeric *KIR*-gene haplotypes commonly observed in Caucasoids, and their linkage disequilibrium in Spanish individuals. Positive and negative relative linkage disequilibrium values are represented with red and gray arrows, respectively. Dotted lines indicate non-statistically significant LD, whilst thickness of solid lines indicates the level of statistical significance ($p$ < 0.05/0.01/0.0001). Conserved "framework" genes are represented as solid boxes. Genes and intergenic spaces are not depicted to scale.

and primer mixes F153/Rt624 (5′–tggtcaggacaarcccctt–3′, exon 3; 5′–aggtccctgcaagggcaa–3′, exon 4) or Fg539/Rc959 (5′–acttctttctgcacaaagagg–3′, exon 4; 5′–cmactcgtagggagagtg–3′, exon 5). PCR conditions were: 1 min at 95°C, then 10 cycles of 30 s at 94°C, 30 s at 64°C and 120 s at 72°C; 20 cycles of 30 s at 94°C, 30 s at 60°C and 120 s at 72°C; final elongation of 10 min at 72°C. Exon sequences were determined using internal primers (not shown). To assess presence of a *KIR3DL1/L2* chimera (*3DL1*060*) (44, 45), its third through fifth and seventh through ninth exons were amplified separately using, respectively, primer mixes Fi2c−201/Ri5+305 (5′–tctagtaagagttgcttctc–3′, intron 2; 5′–atgggcttctgggaaatgga–3′, intron 5); and Fi6g−235/Ra1461 (5′–gagaaagcaggagaaagctg–3′, intron 6; 5′–gttcattggatctggcaacct–3′, exon 9). PCR conditions were: for exons 2–5, 2 min at 95°C; 5 cycles of 30 s at 94°C, 30 s at 60°C and 90 s at 72°C; 25 cycles of 30 s at 94°C, 30 s at 56°C and 90 s at 72°C; and 7 min at 72°C; and for exons 7–9, 2 min at 95°C, 5 cycles of 30 s at 94°C, 30 s at 66°C and 90 s at 72°C; 25 cycles of 30 s at 94°C, 30 s at 62°C and 90 s at 72°C; and 7 min at 72°C. Genotyping was submitted on a regular basis to the external proficiency tests organized by the UCLA Immunogenetics Center (International KIR DNA exchange) to ensure its sensitivity, specificity and consistency, by means of comparison with the results obtained by other labs on samples distributed by the provider.

## Haplotype Assignment

Centromeric and telomeric *KIR*-gene arrangements were inferred in each individual by comparing their gene profile with the common and well-characterized haplotypes shown in **Figure 1**, assuming as few atypical or unknown combinations as possible. In particular, with the exceptions mentioned in the Results section, the following general assumptions were made, based on previous physical mapping and family segregation analyses, and on linkage disequilibrium (LD) between genes, confirmed in our samples using PHASE v2.1 software (46, 47) (results not shown): (i) *KIR 2DP1-2DL1*, and *3DL1-2DS4* are fixed blocks in complete linkage. (ii) Full and deleted *3DP1* variants mark cen-B2 and cen-A/cen-B1 haplotypes, respectively. (iii) Members of pairs *3DL1/3DS1*, and *2DL3/2DL2* behave as alleles of the same locus. (iv) Similarly, *2DS3* and *2DS5* were considered allotypes of a duplicated locus, associating invariably with *2DL5*. (v) The duplicated *2DL5-2DS3/S5* cluster was assigned to the centromeric or the telomeric sides, or both, according to presence or absence of adjacent genes: *2DS2-2DL2* and *2DP1-2DL1* (centromeric); and *3DS1* and *2DS1* (telomeric). (vi) Ambiguities derived from the latter rule were solved by *2DL5* subtyping and taking into account the fixed associations of *2DL5A*001* with *2DS5* and *2DL5A*005* with *2DS3* in tel-B1 and tel-B2 haplotypes, respectively (15, 41); besides those of *2DL5B* with *3DP1*003* and centromeric forms of *2DS3* (or, rarely in

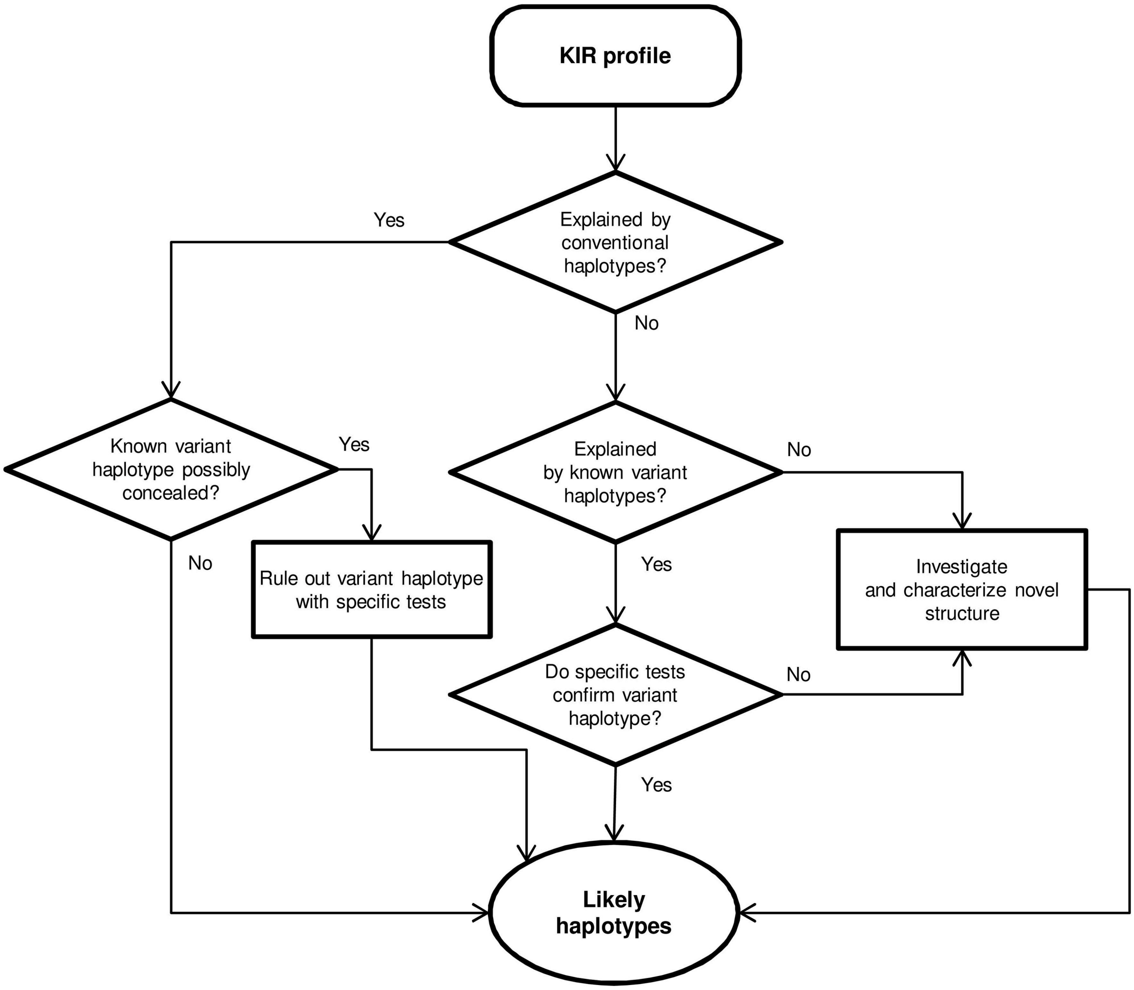

KIR profile
Explained by conventional haplotypes?
Yes
No
Known variant haplotype possibly concealed?
Yes
No
Rule out variant haplotype with specific tests
Explained by known variant haplotypes?
No
Yes
Investigate and characterize novel structure
Do specific tests confirm variant haplotype?
No
Yes
Likely haplotypes

**FIGURE 2 |** Flowchart for haplotype estimation from *KIR*-gene profiles.

Caucasoids, *2DS5*). As shown in **Figure 2**, and detailed in Results, *KIR* gene profiles not fitting with these rules were then compared with contracted and expanded haplotypes described in detail by us and others, and presence of these or new arrangements was verified, when appropriate, by genotyping characteristic traits, such as hybrid genes, or characterized by *de novo* sequencing. Complete haplotypes were assigned only when linkage in cis of centromeric and telomeric motifs was unambiguous; a common ambiguity was presence of two different motifs on both the centromeric and the telomeric intervals, circumstance in which no complete haplotypes were assigned. Relative linkage disequilibrium (D′) between common centromeric and telomeric haplotypes, and its statistical significance were estimated with CubeX (48).

## Characterization of *KIR2DL3**033*

A new *KIR2DS2/2DL3* hybrid was identified by sequencing a partial genomic fragment amplified with primers for exons 5 and 9 (details available upon request). To fully characterize the new hybrid *KIR*, *2DL3**033*, we amplified its complete gene by long-range PCR, using Advantage-2 polymerase mix; forward primer LFc−444 (5′–gctattctgatgcctctggtttagtac−3′), which recognizes a sequence conserved 5′ of most *KIR*, but not in previously known *2DL3* alleles; and the reverse primer LRt1375 (5′–caggagacaactttggatca−3′), specific for a stop codon unique of *2DL3*. PCR conditions were: 2 min at 95°C; 5 cycles of 20 s at 94°C, 30 s at 68°C and 15 min at 72°C; 30 cycles of 20 s at 94°C, 30 s at 64°C; and 15 min at 72°C. The ∼14-Kbp amplicon, spanning from the 5′UT region to the stop codon, was sequenced with internal primers. Confirmatory sequences for the *KIR2DL3**033* stop codon and its new polymorphisms in introns 6 and 7 were obtained from an additional 3.7-Kbp amplicon generated with forward primer Fi6t+1516, (5′–catcctaaagtactgggataact−3′, intron 6) and reverse primer Rt1460 (5′–acattggagctggcaaccca−3′, 3′UT), using the following PCR profile: 2 min at 95°C; 10 cycles of 20 s at 94°C, 30 s at 65°C and 4 min at 72°C; 20 cycles of 20 s at 94°C, 30 s at 61°C; and 4 min at 72°C.

To map *KIR2DL3*033* within the KIR complex, gene walking (26, 41) was carried out – a ∼7-Kbp amplicon spanning exons 7–9 of the preceding gene, the intergenic region and exons 1–4 of the target gene was generated by PCR for 2 min at 95°C; 30 cycles of 20 s at 94°C and 15 min at 72°C; and 20 min at 72°C with a *KIR*-generic forward primer (Fi6–81, 5′–ctaaagagacgttgtatgtggttacc–3′, intron 6) and the gene-specific reverse primer LRa546 (5′–ctccaatgaggtgcaaagtgtccttat–3′, exon 4).

Presence of *KIR2DL3*033* in genomic DNA samples was screened by PCR-SSP, using BioTaq DNA polymerase (Bioline, London, UK) and primer mix Fi6t+1516/Ri6t+2713 (5′–catcctaaagtactgggataact–3′ and 5′–tctgtgctggaggattctga–3′), which recognizes a combination of polymorphisms unique to intron 6 of the *KIR2DS2/2DL3* hybrid, generating a 1387-bp amplicon. A primer pair recognizing a non-polymorphic sequence of the *COCH* gene served as an internal positive control of ∼2 Kbp (COCH-Fi8–86, gaaagaaacttgtgtgttgtctggt; COCH-Ri11+95, attgggtaaagccacaggtgtttg). PCR conditions were: initial denaturation for 2 min at 95°C; 10 cycles of 20 s at 94°C, 30 s at 65°C and 90 s at 72°C; 20 cycles of 20 s at 94°C, 30 s at 61°C and 90 s at 72°C; and 7 min at 72°C.

**TABLE 1 |** Carrier frequencies of *KIR* genes, pseudogenes, and their main structural and positional variants in 414 Spanish donors.

| Long tailed | | | Short tailed | | | Others | | |
|---|---|---|---|---|---|---|---|---|
| **Gene** | **Variant** | **%** | **Gene** | **Variant** | **%** | **Gene** | **Variant** | **%** |
| *2DL1* | | 96.6 | *2DS1* | | 42.8 | *2DP1* | | 96.6 |
| *2DL2* | | 58.0 | *2DS2* | | 58.7 | *3DP1* | *All* | 100.0 |
| *2DL3* | | 88.4 | *2DS3* | *All* | 34.8 | | *Exon 2*$^{+}$ | 31.4 |
| *2DL4* | | 100.0 | | *Centromeric* | 31.2 | | *Exon 2*$^{del}$ | 96.6 |
| *2DL5* | *All* | 57.7 | | *Telomeric* | 15.2 | *3DX1* | | 100.0 |
| | *Centromeric* | 32.1 | *2DS4* | *All* | 95.9 | | | |
| | *Telomeric* | 41.5 | | *Correct CDS* | 36.2 | | | |
| *3DL1* | | 96.1 | | *Frame-shifted* | 82.1 | | | |
| *3DL2* | | 100.0 | *2DS5* | *All* | 30.0 | | | |
| *3DL3* | | 100.0 | | *Centromeric* | 0.0 | | | |
| | | | | *Telomeric* | 30.0 | | | |
| | | | *3DS1* | | 42.5 | | | |

*CDS, coding sequence.*

## Genomic Characterization of *KIR2DL5B*0020202*

The complete coding region and part of the intervening introns of the new allele *KIR2DL5B*0020202* were derived from a ∼9.4-Kbp fragment generated from donor D139 by long range PCR with primer mix Fg–97b/Rg1769b (5′–tcaccctcccrtgatgtg–3′, promoter region; and 5′–ggaaggtggaacagcacgtgtctc–3′, 3′UTR) and Advantage-2 polymerase. PCR conditions were: 2 min at 95°C; 30 cycles of 20 s at 94°C and 15 min at 72°C; and 20 min at 72°C. The relative position of this allele in the *KIR* complex was determined by gene walking (26, 41). Identical procedures were used in another donor to identify and map *KIR2DL5B*0020106* (32).

## DNA Sequencing and Nomenclature

PCR products were submitted, with no cloning step, to direct nucleotide sequencing in both strands, using internal primers (sequences available upon request). The products were analyzed in an ABI Prism 3100-Avant Genetic analyzer (Applied Biosystems) in the central DNA sequencing facility of *Instituto de Investigación Sanitaria Puerta de Hierro Segovia de Arana (IDIPHISA)*. The names *KIR2DL3*033* and *KIR2DL5B*0020202* (EMBL/GenBank/DDBJ database accession numbers HG931348 and LT604077, respectively) were officially assigned by the WHO Nomenclature Committee for factors of the HLA System, Subcommittee for Killer-cell Immunoglobulin-like Receptors (13).

## Flow Cytometry

The NK-cell population was defined in PBMCs by the CD3$^{-}$CD56$^{+}$ phenotype, using anti-CD3-VioBlue (Miltenyi Biotec, Bergisch Gladbach, Germany) and anti-CD56-APC (eBioscience, Inc, San Diego, CA). These were combined in a donor carrying *KIR2DL3*033*, with anti-KIR2DL3-FITC (180701, R&D Systems, Minneapolis, MN, USA) and anti-KIR2DL3/L2/S2-PE/Cy7 (DX27, Miltenyi Biotec); and in donors bearing transcribed *KIR2DL5B*002* alleles, with anti-KIR2DL5-PE [UP-R1 (49), Biolegend, San Diego, CA, USA]. Isotype-matched negative controls were IgG1-PE (clone MOPC-21, Sigma-Aldrich, St. Louis, MO), IgG2a-FITC (clone MG2A01, Invitrogen, Camarillo, CA, USA), and IgG2a-PE (clone S43.10, Miltenyi Biotec). Flow cytometry analyses were performed in a MACSQuant Analyzer using MACSQuantify software (both by Miltenyi Biotec) in the central facility of *IDIPHISA*.

# RESULTS

## *KIR*-Gene Frequencies and Profiles

To characterize and understand the diversity of *KIR* genotypes in the Spanish population, we determined the *KIR*-gene content in the genome of 414 unrelated donors using a locally designed PCR-SSP (sequence-specific primers) method, which also discriminates between structural variants of *KIR 2DS4* and *3DP1*. Unusual genotypes were further investigated and confirmed using a combination of techniques, including, as appropriate, probe hybridization; selective sequencing of the relevant coding, non-coding or intergenic regions; and physical mapping of neighbor genes (*KIR*-gene "walking"), when relevant.

Individual *KIR*-gene frequencies are shown in **Table 1**. Framework *KIR* genes and pseudogenes *3DL3*, *3DP1*, *2DL4,* and *3DL2* were detected in every donor, even though seven of them were then deduced to lack *3DP1-2DL4-3DL1/S1*, or have *3DL2* partially deleted, on one chromosome. Also found in all 414 donors was *KIR3DX1*, a gene of uncertain function located outside and 180 Kbp centromeric to the *KIR* complex, which is seldom typed for. Of note, only the latter gene and *KIR3DL3*, both of unknown biological significance (5, 50), appear to be truly conserved (i.e., not submitted to CNV) among human *KIR*.

Non-framework genes encoding long-tailed KIR typical of A-haplotypes *2DL1*, *2DL3,* and *3DL1* had frequencies of ∼90%; whilst *2DL2* and *2DL5*, encoding inhibitory KIR characteristic of B-haplotypes were seen in ca. 60% of the donors. Activating *KIR*-gene frequencies were more variable, ranging from 30.0% (*2DS5*) to 95.9% (*2DS4*). The latter was most often represented by frame-shifted alleles (82.1 vs. 36.2% alleles with canonical coding sequence), therefore the most common functional activating *KIR* was actually *2DS2* (58.7%) (**Table 1**).

Based on *KIR*-gene content, we found 38 different profiles, which were sorted into three groups: (i) the AA profile carrying exclusively genes of the A haplotype (24.15% of individuals); (ii) 16 BX profiles having all genes of the A-haplotype plus one or more B-haplotype genes (61.12%); and (iii), 21 BB profiles, defined by presence of B-haplotype genes and lack of one or more genes of the A-haplotype (14.73% of individuals). The individual and grouped frequencies of those profiles are shown in **Figures 3A,B**. The AA genotype was most frequent (ID: 1, 24.15%), followed by six BX genotypes (ID: 4, 2, 5, 7, 3 and 6) which, together, account for more than 50% of individuals; and by the two most common BB profiles (ID: 71 and 72; 3.14 and 2.66%, respectively). The overall distribution of *KIR*-gene frequencies and profiles observed in our sample is not dissimilar from those reported in other Caucasoid populations, and it is also consistent with those found in other large samples of Spanish individuals (40, 51–53).

A cumulative frequency analysis of the distribution of *KIR* genotypes is shown in **Figure 3C**. More than 50% of the population can be represented by three genotypes; six genotypes account for 75% of the total; and 13 are needed to explain 90% of the diversity. The remaining 10% is accounted for by 25 gene profiles, 14 of which were observed only once. As analyzed in more detail below, more than one third of the gene profiles (fourteen genotypes, belonging to 20 individuals) cannot be explained by any homo- or heterozygous combination of canonical haplotypes.

## Centromeric and Telomeric Interval Analysis

Following a parsimonious approach that assumed as few non-canonical or novel *KIR*-gene arrangements as possible, we assigned each gene profile to the most likely diploid combination of well-known centromeric and telomeric haplotypes seen in most ethnicities (**Figure 1**). This was possible in 394 donors (95.17%), and the frequencies of their centromeric and telomeric deduced genotypes are represented in **Figure 4A**. On the centromeric region, the cen-AA genotype was most common (40.34%), followed by combinations of cen-A with either cen-B1 or cen-B2 haplotypes, seen with even frequencies−21.23 and 22.95%, respectively. Much less frequent were combinations of only cen-B1 or -B2 motifs, each seen at <5%. Centromeric motifs consistently associated with the major *KIR3DP1* allotypes (exon $2^{+/\mathrm{del}}$) as reported (54), with a single exception described previously by us [haplotype *2DS2-2DL2-3DP1*$^{del}$ in family C180 (43)]. On the telomeric side, the combination of two tel-A segments was again most common at 55.07%; followed by combinations of tel-A and tel-B motifs (24.88% tel-AB1 and 11.35% tel-AB2); whereas combinations of tel-B1 or -B2 haplotypes collectively represented only 3.83% of telomeric genotypes.

From the previous genotypes, and from those eventually assigned to individuals with unconventional profiles (following section), we estimated the frequencies of individual centromeric and telomeric haplotypes (**Figure 4B**). Also based on those assignments of partial centromeric and telomeric motifs, entire basic haplotypes (i.e., A vs. B) could be phased in 81% of donors (i.e., those who, on at least one of their telomeric or centromeric segments, only had either A or B profiles, but not both). This allowed a minimum estimate of the frequency of such complete basic haplotypes (**Figure 4C**): at least 45% of haplotypes are A (cen-A/tel-A), whilst more than 36% of haplotypes would be of the B-type (cen-A/tel-B, cen-B/tel-A, or cen-B/tel-B). The remaining ∼19% of haplotypes could not be unambiguously assigned to the A- or the B-groups, because coincidence of A and B profiles on both the centromeric and the telomeric regions of the same individual precluded phasing.

Linkage disequilibrium between common centromeric and telomeric motifs is analyzed in **Table 2**, and depicted in **Figure 1**. Of note, association between cen-A and tel-A haplotypes was weak (D' = 0.089) and non-significant, suggesting that their common composition of a complete A haplotype is explained merely by the high frequency of its two segments. In contrast, strong were the positive LD of tel-B2 (*KIR 3DS1-2DL5-2DS3-2DS1*) with cen-B1, and its negative LD with cen-A (0.70 and −0.73, respectively; $p < 0.0001$). As previously noted (17), this means that the telomeric cluster *2DL5A*005-2DS3*002* is most often associated with a nearly identical sequence (*2DL5B*002-2DS3*001*) on the centromeric segment of the same haplotype, perhaps a reminiscence of their recent common origin.

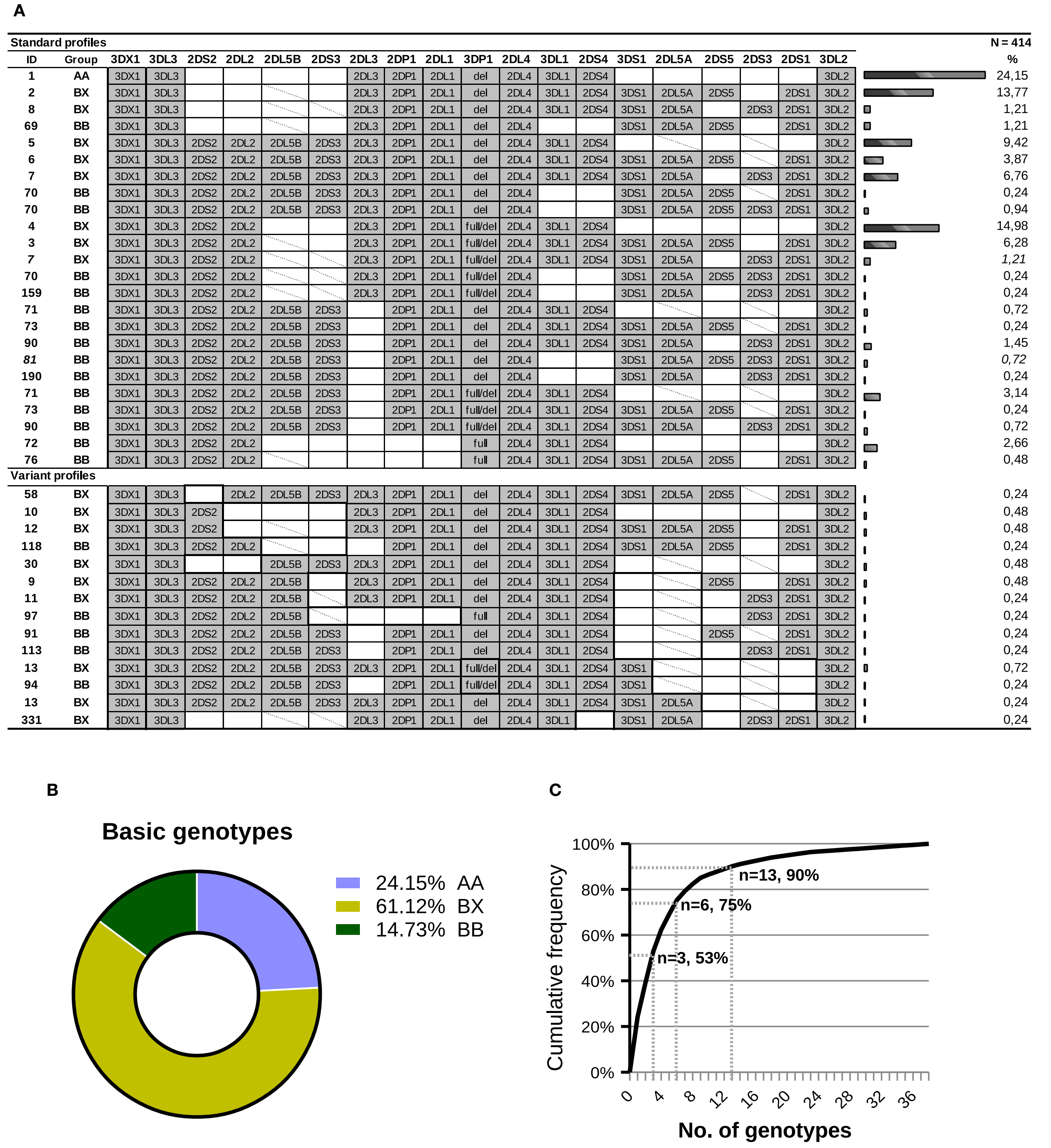
Taken together, only 24 of the 38 observed *KIR*-gene profiles were explained by combinations of canonical centromeric and telomeric motifs; whilst no such combination could account for 14 profiles (i.e., more than one third), which owed to 2.66% of all chromosomes carrying atypical *KIR*-gene arrangements, as analyzed in the following sections.

## *KIR* Gene Profiles Explained by Known Non-canonical Arrangements

Twenty individuals (4.8%) had *KIR* genotypes unexplained by any homo- or heterozygous combination of conventional haplotypes. Of those, 17 could be explained, as detailed below, by recombinant or variant structures previously described by us and others. Furthermore, specific tests targeting marker polymorphisms disclosed two additional donors in whom atypical haplotypes were concealed under apparently canonical *KIR* profiles, making a total of 22 individuals (5.3%) with unusual *KIR*-gene arrangements (**Figure 5**). Finally, three donors carried new haplotypes or hybrid genes characterized in the following sections.

### Centromeric Interval

Two atypical profiles with singularities affecting only the centromeric region were found in seven donors. One of them



A

| Standard profiles | | | | | | | | | | | | | | | | | | | | | N = 414 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Group | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | 3DP1 | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | 2DS3 | 2DS1 | 3DL2 | % |
| 1 | AA | 3DX1 | 3DL3 | | | | | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | | | | | 3DL2 | 24,15 |
| 2 | BX | 3DX1 | 3DL3 | | | ╲ | | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 13,77 |
| 8 | BX | 3DX1 | 3DL3 | | | ╲ | ╲ | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 1,21 |
| 69 | BB | 3DX1 | 3DL3 | | | ╲ | | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 1,21 |
| 5 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | | ╲ | | 3DL2 | 9,42 |
| 6 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | ╲ | 2DS1 | 3DL2 | 3,87 |
| 7 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 6,76 |
| 70 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | ╲ | 2DS1 | 3DL2 | 0,24 |
| 70 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | 2DS3 | 2DS1 | 3DL2 | 0,94 |
| 4 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | | | 2DL3 | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | | | | | | 3DL2 | 14,98 |
| 3 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | ╲ | | 2DL3 | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 6,28 |
| *7* | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | ╲ | ╲ | 2DL3 | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | *1,21* |
| 70 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | ╲ | ╲ | 2DL3 | 2DP1 | 2DL1 | full/del | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | 2DS3 | 2DS1 | 3DL2 | 0,24 |
| 159 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | ╲ | ╲ | 2DL3 | 2DP1 | 2DL1 | full/del | 2DL4 | | | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 0,24 |
| 71 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | | ╲ | | 3DL2 | 0,72 |
| 73 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | ╲ | 2DS1 | 3DL2 | 0,24 |
| 90 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 1,45 |
| *81* | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | 2DS3 | 2DS1 | 3DL2 | *0,72* |
| 190 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | | | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 0,24 |
| 71 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | | ╲ | | ╲ | | 3DL2 | 3,14 |
| 73 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | ╲ | 2DS1 | 3DL2 | 0,24 |
| 90 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 0,72 |
| 72 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | | | | | | full | 2DL4 | 3DL1 | 2DS4 | | | | | | 3DL2 | 2,66 |
| 76 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | ╲ | | | | | full | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 0,48 |
| Variant profiles | | | | | | | | | | | | | | | | | | | | | |
| 58 | BX | 3DX1 | 3DL3 | | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | ╲ | 2DS1 | 3DL2 | 0,24 |
| 10 | BX | 3DX1 | 3DL3 | 2DS2 | | | | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | | | | | 3DL2 | 0,48 |
| 12 | BX | 3DX1 | 3DL3 | 2DS2 | | ╲ | | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 0,48 |
| 118 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | ╲ | | | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 0,24 |
| 30 | BX | 3DX1 | 3DL3 | | | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | | ╲ | | 3DL2 | 0,48 |
| 9 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | 2DS5 | | 2DS1 | 3DL2 | 0,48 |
| 11 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | ╲ | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | | 2DS3 | 2DS1 | 3DL2 | 0,24 |
| 97 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | ╲ | | | | full | 2DL4 | 3DL1 | 2DS4 | | ╲ | | 2DS3 | 2DS1 | 3DL2 | 0,24 |
| 91 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | 2DS5 | ╲ | 2DS1 | 3DL2 | 0,24 |
| 113 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | | ╲ | | 2DS3 | 2DS1 | 3DL2 | 0,24 |
| 13 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | ╲ | | ╲ | | 3DL2 | 0,72 |
| 94 | BB | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | full/del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | ╲ | | ╲ | | 3DL2 | 0,24 |
| 13 | BX | 3DX1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | ╲ | | 3DL2 | 0,24 |
| 331 | BX | 3DX1 | 3DL3 | | | ╲ | ╲ | 2DL3 | 2DP1 | 2DL1 | del | 2DL4 | 3DL1 | | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 0,24 |

**FIGURE 3 |** *KIR*-gene profiles observed in a sample of Spanish individuals. **(A)** Gene presence or absence is represented by solid gray and empty boxes; allelic forms are indicated for *KIR3DP1*. Carrier frequency is given on the right side of each genotype. Gene order reflects, approximately, that seen in the *KIR* complex, with genes forming A-haplotypes in the middle, flanked on both sides by genes characteristic of B-haplotypes. For *KIR 2DL5* and *2DS3*, genes represented by two paralogues, a diagonal line indicates that the gene is present in a genotype, but most likely in the alternative location. The ID column shows the number by which the genotype is registered in www.allelefrequencies.net (40); since this database does not distinguish between structural/positional variants, a same ID can correspond to several profiles in the table. Two profiles (7 and 81) include each one individual bearing a concealed variant haplotype, as explained in the text. In the lower part of the panel, which compiles gene profiles not explained by conventional haplotypes, thick lines highlight distinctive traits, including missing genes normally associated with ones present in a given genotype. **(B)** Distribution of the major groups of *KIR* genotypes. **(C)** Cumulative frequencies of *KIR* profiles.

A

| Centromeric | | | | | | | | | | N = 414 |
|---|---|---|---|---|---|---|---|---|---|---|
| Genotype | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | 3DP1 | % |
| AA | 3DL3 | | | | | 2DL3 | 2DP1 | 2DL1 | del | 40.34 |
| AB1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | 2DL3 | 2DP1 | 2DL1 | del | 21.23 |
| AB2 | 3DL3 | 2DS2 | 2DL2 | | | 2DL3 | 2DP1 | 2DL1 | full/del | 22.95 |
| B1B1 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | del | 3.37 |
| B1B2 | 3DL3 | 2DS2 | 2DL2 | 2DL5B | 2DS3 | | 2DP1 | 2DL1 | full/del | 4.10 |
| B2B2 | 3DL3 | 2DS2 | 2DL2 | | | | | | full | 3.14 |

| Telomeric | | | | | | | | | | N = 414 |
|---|---|---|---|---|---|---|---|---|---|---|
| Genotype | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | 2DS3 | 2DS1 | 3DL2 | % |
| AA | 2DL4 | 3DL1 | 2DS4 | | | | | | 3DL2 | 55.07 |
| AB1 | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 24.88 |
| AB2 | 2DL4 | 3DL1 | 2DS4 | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 11.35 |
| B1B1 | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | | 2DS1 | 3DL2 | 1.45 |
| B1B2 | 2DL4 | | | 3DS1 | 2DL5A | 2DS5 | 2DS3 | 2DS1 | 3DL2 | 1.90 |
| B2B2 | 2DL4 | | | 3DS1 | 2DL5A | | 2DS3 | 2DS1 | 3DL2 | 0.48 |

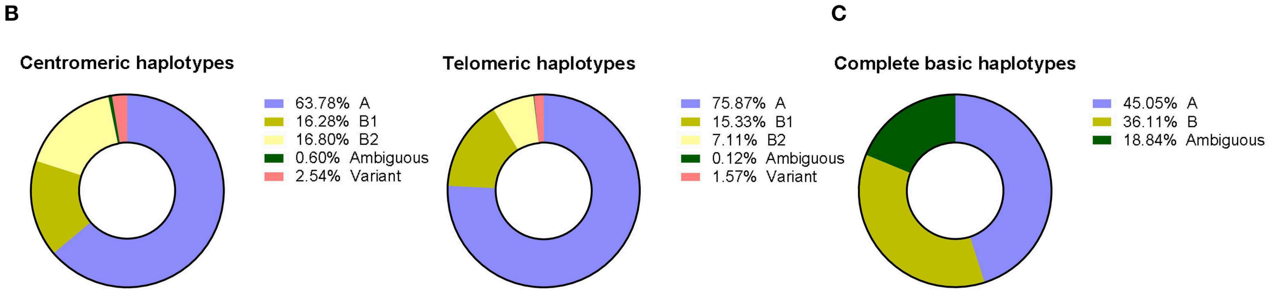


**FIGURE 4 |** Distribution of centromeric and telomeric gene profiles and haplotypes. **(A)** Centromeric and telomeric profiles corresponding to combinations of conventional haplotypes are represented as in **Figure 2**. Profiles derived from variant and novel arrangements are not represent, therefore frequencies do not sum up 100%. **(B)** Distribution of partial *KIR* haplotypes. **(C)** Frequencies of complete haplotypes and phasing ambiguities.

**TABLE 2 |** Linkage disequilibrium between centromeric and telomeric *KIR* haplotypes.

| hap. freq.<br>D′<br>p | tel-A | tel-B1 | tel-B2 |
|---|---|---|---|
| cen-A | 0.503 | 0.129 | 0.012 |
| | 0.089 | **0.546** | **−0.727** |
| | n.s. | <0.05 | <0.0001 |
| cen-B1 | 0.096 | 0.012 | 0.054 |
| | **−0.253** | −0.534 | **0.702** |
| | <0.01 | n.s. | <0.0001 |
| cen-B2 | 0.155 | 0.011 | 0.003 |
| | **0.692** | −0.574 | −0.775 |
| | <0.05 | n.s. | n.s. |

*N.s., non-significant. Bold, statistically significant D′ values.*

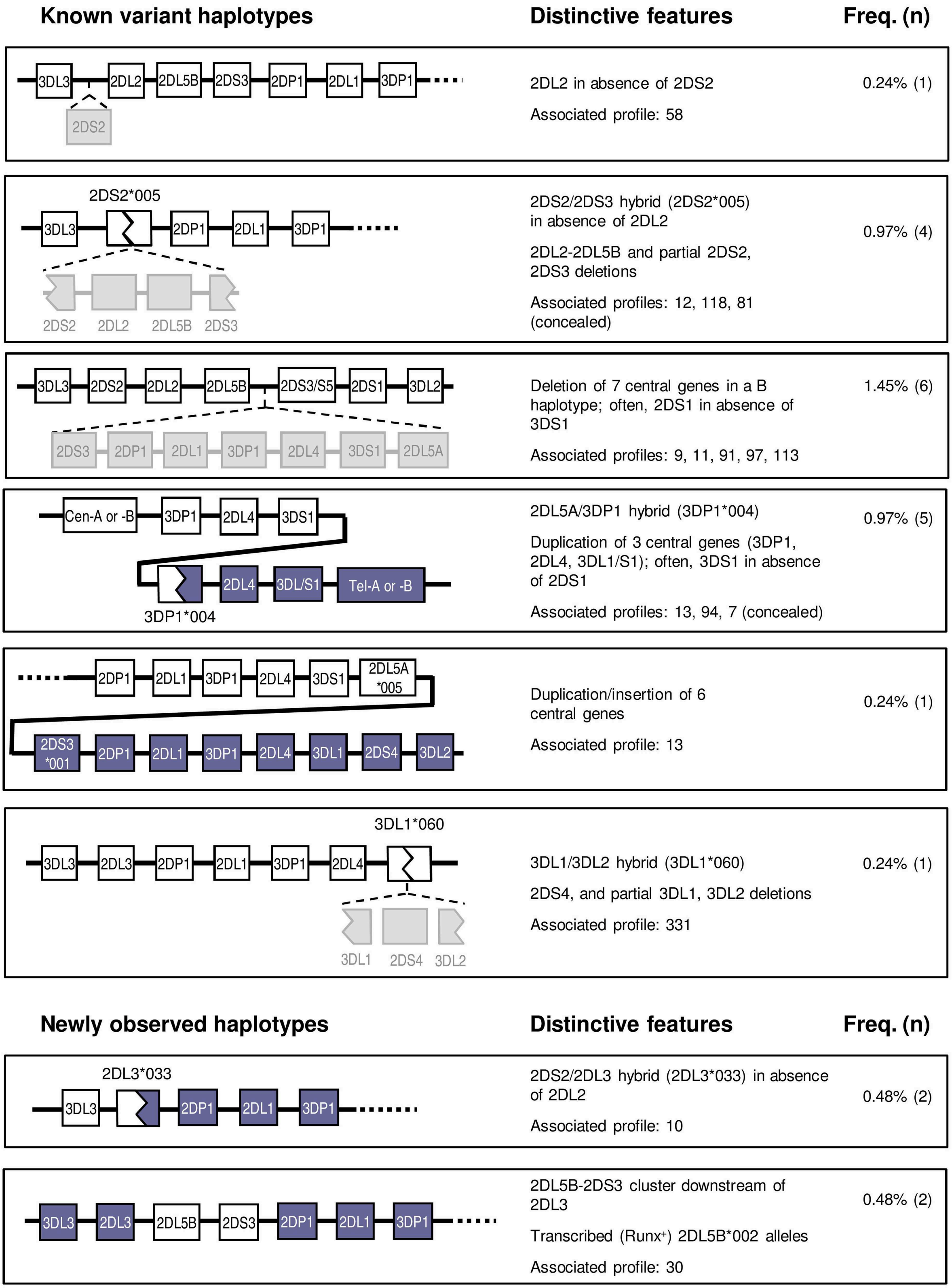
presented *KIR2DL2* in absence of *2DS2*. This profile, initially characterized in Black individuals (8, 19), might be explained by *KIR2DS2* deletion from the usual *2DS2-2DL2* cluster; alternatively, since these two highly homologous genes possibly derive from a common ancestor, the variant profile might be reminiscent of the ancestral haplotype existing before the gene duplication that gave origin to the two paralogues (**Figure 5**).

Another five individuals had the reciprocal combination, i.e., *KIR2DS2* in isolation from *2DL2* and/or other cen-B1 characteristic genes, a profile typically seen in a shortened haplotype previously characterized by our group (43). Such haplotype, generated by unequal crossing-over between *KIR2DS2* and *2DS3*, is marked by the resulting chimeric allele, *KIR2DS2***005*, and by deletion of the intervening genes (*KIR2DL2*, *2DL5B*, and *2DS3*). A PCR-SSP test specific for *KIR2DS2***005* confirmed this marker in three of the five suspected individuals; and disclosed it in another donor with an apparently standard *KIR* profile, in which lack of *KIR2DL2* within the recombinant haplotype was concealed by its presence on the other one.

The remaining two *KIR2DS2*$^{+ve}$*-2DL2*$^{-ve}$ individuals did not carry this shortened B haplotype, since they were negative in the *KIR2DS2***005* test. Instead, they carried a new recombinant haplotype and a novel hybrid gene, as analyzed in a separate section. Also described separately is the last variant centromeric profile, in which the *KIR 2DL5B-2DS3* cluster is seen in absence of its usual companions *2DS2-2DL2.*

## Central Cluster

Affecting both the centromeric and the telomeric regions were haplotype arrangements that included duplications or deletions of the central framework genes, identified in 13 donors. For instance, an atypical profile marked by presence of *KIR 2DS3/2DS5* and *2DS1* in absence of *3DS1* was explained by previously characterized haplotypes bearing an extensive deletion affecting seven genes on the central region of the *KIR*-gene complex, including, among others, the otherwise conserved *KIR 3DP1*, *2DL4*, and *3DL1/3DS1* loci (19, 20). The resulting contracted B haplotype, containing only seven genes, generates an unusual juxtaposition of the centromeric *KIR2DL5B* and the telomeric *KIR2DS3/2DS5* loci. This profile was observed in six individuals, but in contrast with studies on other ethnic groups, the deletion of the central genes was never observed in homozygosis in this population sample. However, such genotype does exist among Spanish Caucasoids, as we reported earlier [donor LH304 (49)].



| Known variant haplotypes | Distinctive features | Freq. (n) |
|---|---|---|
| | 2DL2 in absence of 2DS2<br>Associated profile: 58 | 0.24% (1) |
| | 2DS2/2DS3 hybrid (2DS2*005) in absence of 2DL2<br>2DL2-2DL5B and partial 2DS2, 2DS3 deletions<br>Associated profiles: 12, 118, 81 (concealed) | 0.97% (4) |
| | Deletion of 7 central genes in a B haplotype; often, 2DS1 in absence of 3DS1<br>Associated profiles: 9, 11, 91, 97, 113 | 1.45% (6) |
| | 2DL5A/3DP1 hybrid (3DP1*004)<br>Duplication of 3 central genes (3DP1, 2DL4, 3DL1/S1); often, 3DS1 in absence of 2DS1<br>Associated profiles: 13, 94, 7 (concealed) | 0.97% (5) |
| | Duplication/insertion of 6 central genes<br>Associated profile: 13 | 0.24% (1) |
| | 3DL1/3DL2 hybrid (3DL1*060)<br>2DS4, and partial 3DL1, 3DL2 deletions<br>Associated profile: 331 | 0.24% (1) |

| Newly observed haplotypes | Distinctive features | Freq. (n) |
|---|---|---|
| | 2DS2/2DL3 hybrid (2DL3*033) in absence of 2DL2<br>Associated profile: 10 | 0.48% (2) |
| | 2DL5B-2DS3 cluster downstream of 2DL3<br>Transcribed (Runx+) 2DL5B*002 alleles<br>Associated profile: 30 | 0.48% (2) |

**FIGURE 5 |** Variant and novel haplotypes detected in Spanish individuals. Colors are used to highlight gene deletions and hybrid structures derived from recombination of two different genes or haplotypes, but they lack a specific meaning. Duplicated genes are represented in parallel.

Five individuals showed the opposite combination—*KIR3DS1* in absence of *2DS1* (and the telomeric *2DL5-2DS5/S3* group). This trait is typically seen associated with a tandem duplication of the *3DP1-2DL4-3DL1/S1* cluster and the recombinant, transcribed *KIR3DP1**004* allotype, a marker of this duplication (21, 22). This "full" (exon $2^{+ve}$) *3DP1* allotype is often discordant at first glance with the centromeric profile, serving as a beacon of the expanded haplotype. Its presence in four of the five suspected individuals, as well as in one with an apparently normal genotype, was confirmed by specific PCR-SSP detection of the hybrid gene *KIR3DP1**004*.

A sixth donor with a similar *KIR* genotype (*3DS1*$^{+ve}$-*2DL5*$^{+ve}$-*2DS3*$^{+ve}$-*2DS1*$^{-ve}$) was, in contrast, negative for *3DP1**004*. Its atypical gene profile could instead be explained by presence of an expanded haplotype we described in one Irish Caucasoid (17). Such haplotype is characterized by an even larger duplication/insertion of six central *KIR* genes, due to unequal crossing-over between the intergenic regions of the telomeric and the centromeric *2DL5-2DS3* clusters (**Figure 5**). Consistently with such duplication, the donor had two *3DL1* alleles besides *3DS1* (not shown).

### Telomeric Interval

Least common were variants affecting exclusively the telomeric *KIR* region—we observed a single unusual *3DL1*$^{+ve}$-*2DS4*$^{-ve}$ profile, represented in one donor. In this individual, we identified another previously described hybrid gene, *KIR3DL1**060*, arising from an asymmetric recombination event that fused *3DL1* exons 1–5 with *3DL2* exons 6–9 (44, 45). Concomitant deletion of the intervening gene *KIR2DS4* explains this gene profile.

## A Novel *KIR2DS2/KIR2DL3* Fusion Gene Within a Recombinant B/A Motif Challenges Conventional Gene and Haplotype Classification

One donor apparently presented *KIR2DS2* in isolation from other B-haplotype genes (i.e., in the context of an AA genotype), but, as stated earlier, was negative for *2DS2**005*, which marks the only known gene arrangement that could account for such profile (43). To characterize the putative *2DS2* gene in the novel gene profile, we amplified its exons 5 through 7 (encoding the D2 Ig-like domain, stem and transmembrane region) in a 7.6-Kbp genomic fragment. Analysis of this amplicon revealed, instead of a proper *KIR2DS2* gene, a new hybrid sequence, of which the 5′- and the 3′-ends were homologous to *2DS2* and *2DL3*, respectively. Based on this information, the new gene was then amplified in a single ∼14-Kbp genomic fragment comprising all its exons and introns, and sequence analysis confirmed its hybrid nature—its 5′ side (down to nucleotide 2,075 of intron 6, ca. 11 Kbp) was identical to *2DS2**0010102*; whilst the rest of the gene (∼3 Kbp through the stop codon) matched *2DL3**0020101,* except for three base substitutions: two unique in intron 6 (11586 A>G and 11849 C>T), and one in intron 7, shared with *2DL3**010* (13627 G>A). The hybrid sequence is likely the result of asymmetric (non-allelic) homologous recombination between *KIR2DS2* and *KIR2DL3* genes present in different B and A haplotypes, respectively. The apparent recombination spot is few bases upstream of an AluSX element (not shown).

As discussed later, the novel structure challenged unwritten rules followed previously to designate chimeric KIR, not sitting comfortably within any of the current designations. It was officially assigned to the *KIR2DL3* locus with the name *2DL3**033* by the KIR Nomenclature Committee, which reflects appropriately the fact that the encoded protein should have an intracellular inhibitory tail identical to those of most other *2DL3* alleles.

To map the new hybrid gene, we used a *KIR*-gene walking approach (i.e., amplification and sequencing of a genomic fragment spanning part of the gene of interest, and one in its vicinity), which showed *2DL3**033* to be located 3′ of *3DL3*, like common alleles of both its homologs *2DS2* and *2DL3* in conventional *KIR* haplotypes. Furthermore, the donor had *3DL3**003*, allele commonly linked to *2DS2* (24), in consonance with presence of a *2DS2*-like 5′ region in *2DL3**033*.

PBMCs of the donors in whom *KIR2DL3**033* was originally found were, unfortunately, unavailable. To estimate the *KIR2DL3**033* distribution and enable expression studies on this allele, we designed a PCR-SSP method targeting a specific combination of polymorphisms in its sixth intron, which led us to identify two additional examples of this allele in 1,101 DNA samples (∼0.2%; confidence interval with $p < 0.05$: 0.00–0.46%). Using PBMC of one of those donors, we could readily amplify the *KIR2DL3**033* complete coding region (∼1.1 kb) by RT-PCR, demonstrating the normal transcription and processing of its mRNA. According to this, *KIR2DL3**033* encodes ligand-recognition Ig-like domains, and a stem homologous to those of the activating 2DS2, but transmembrane and long intracytoplasmic regions like those of 2DL3 (**Figure 6A**). Therefore, the encoded receptor should combine the weak HLA-C1 recognition of the former, and the inhibitory capacity of the latter.

Since PBMCs of *KIR2DL3**033* donors with cenAA profiles were unavailable, we performed multicolour flow cytometry analyses of PBMCs of one available individual encoding KIR2DL3*033 on one haplotype and the 2DS2-2DL2 combination on the second one. The observed lack of staining with a KIR2DL3-specific mAb (**Figure 6B**) was consistent with the 2DS2-like nature of the KIR2DL3*033 ectodomain. Unfortunately, the donor genotype precluded positive identification of cells expressing KIR2DL3*033, since these are indistinguishable with the available mAbs from those bearing its homologues 2DS2 and 2DL2. Flow cytometry studies of further donors with favorable (i.e., cenAA) *KIR* profiles are warranted to demonstrate positively KIR2DL3*033 NK-cell surface expression; this would be marked by presence of cells staining with DX27 (or equivalent mAbs), but not with reagents monospecific for conventional KIR2DL3 alleles.

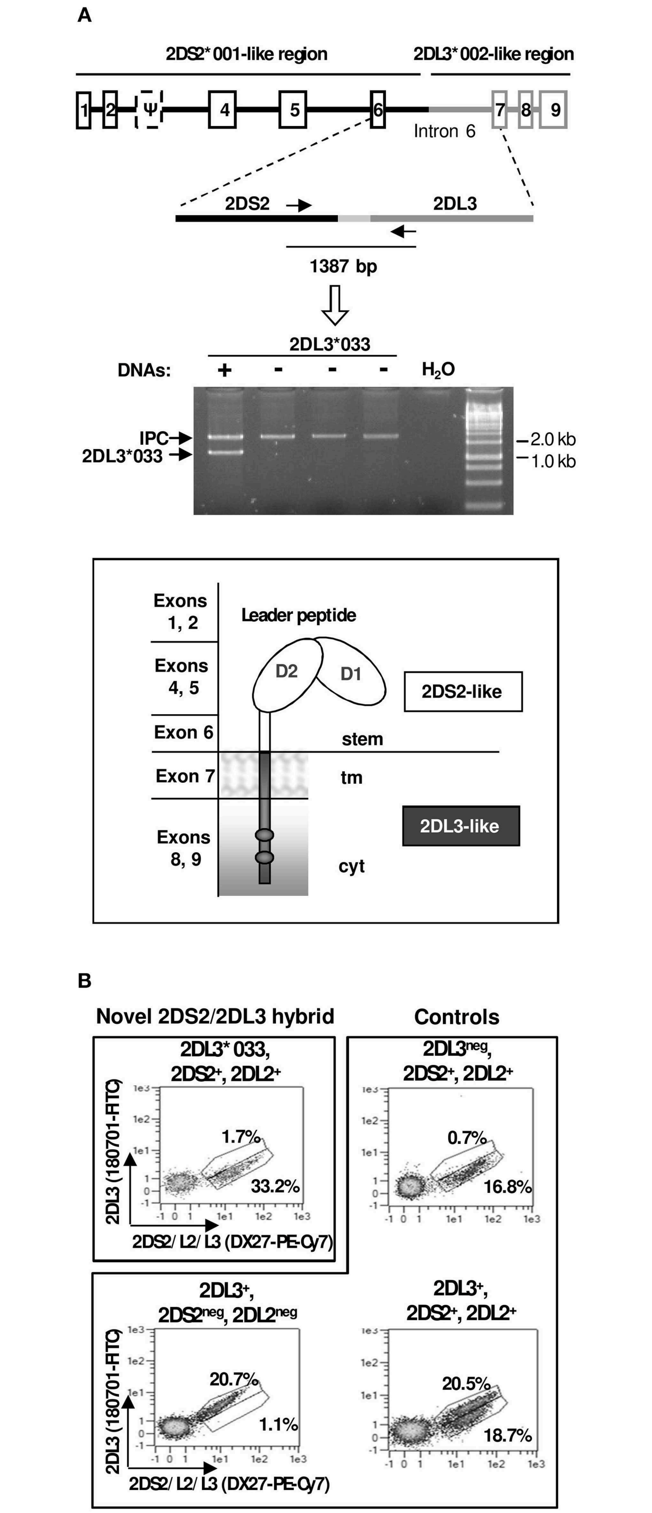


**FIGURE 6 |** Characterization of *KIR2DL3*033*. **(A)** Gene and protein structure, including homology to other KIR, and an example of the PCR-SSP test to identify the novel allele (lane 1). IPC stands for internal positive control. **(B)** Flow cytometry plots of NK cells (CD3$^-$CD56$^+$) from a donor expressing *KIR2DL3*033*, and from others with common *KIR 2DL2/2DS2/2DL3* genotypes.

## *KIR2DL5B*0020202*—A Transcribed Allele Linked to *KIR2DL3* in an Unusual Centromeric B Motif

Two individuals exhibited unusual presence of the *KIR 2DL5-2DS3* genes in absence of any other genes characteristic of *B* haplotypes, to which they are normally linked (i.e., *2DL2* or *3DS1*). To understand the unusual genotype, we used *KIR*-gene walking—amplification of a ∼4-Kbp region spanning *KIR2DL5* exons 1–3, the last (ninth) exon of the unknown preceding gene, and the intergenic region. This enabled us to map, in both donors, *KIR2DL5B*002*-related sequences downstream of *KIR2DL3*010*. This arrangement is seemingly identical to one previously described in a minority of Black Africans (8, 9, 32, 55), in whom a *2DL5-2DS3* block is inserted between *2DL3* and *2DP1-2DL1* (**Figure 5**), thus converting a centromeric A-motif into a B-haplotype, according to the agreed definition of these.

To further characterize *KIR2DL5B* in the unusual haplotype, we amplified the whole gene by long-range PCR (∼9 Kbp). Sequence analysis of this amplicon, along with that derived from *KIR*-gene walking, revealed, in one donor, a new *KIR2DL5B* allele, designated *KIR2DL5B*0020202*, in which a coding region identical to that of the previously known *KIR2DL5B*0020201* is fused to a "*KIR2DL5*-type III" promoter (15), similar to that found in the expressed allele *KIR2DL5B*003* (**Figure 7A**). This type of promoter retains an intact RUNX-binding site seen in all clonally expressed *KIR*, in contrast with common *KIR2DL5B* alleles, in which the site is destroyed by substitution of adenosine for guanosine −97, polymorphism completely associated with epigenetic silencing (41). Similar results were obtained in the second donor, in whom we found allele *KIR2DL5B*0020106*. This allele and the novel *KIR2DL5B*0020202* share a nearly identical combination of promoter and coding sequences, the latter differing by a single synonymous substitution (**Figure 7B**).

Conservation of the RUNX site in *KIR2DL5B*0020106* and **0020202* should, according to our hypothesis (41), confer these alleles a capacity to be transcribed, in contrast with most *2DL5B* alleles. To test this prediction, we performed RT-PCR experiments on RNA isolated from PBMC of donors D139 and 140016. Specific, correctly spliced amplicons were readily obtained for both *KIR2DL5B*0020106* and *KIR2DL5B*0020202* (**Figure 7B**), as verified by direct sequencing. This positive result opened the possibility of KIR2DL5B being expressed on the cell surface, which has never been shown. To explore this, we undertook flow-cytometry assays with the KIR2DL5-specific mAb UP-R1, but these showed no specific surface staining of peripheral blood NK cells (**Figure 7C**). This result is in line with the described behavior of allele KIR2DL5A*005, which shares with KIR2DL5B*002 the identical sequence in the mature protein. Such protein, seemingly due to substitution of Ser for Gly174, is retained intracellularly in NK cells, besides reacting weakly with mAb UP-R1 (56); unfortunately, monoclonal antibodies for intracellular KIR2DL5A*005/B*002 staining are unavailable.

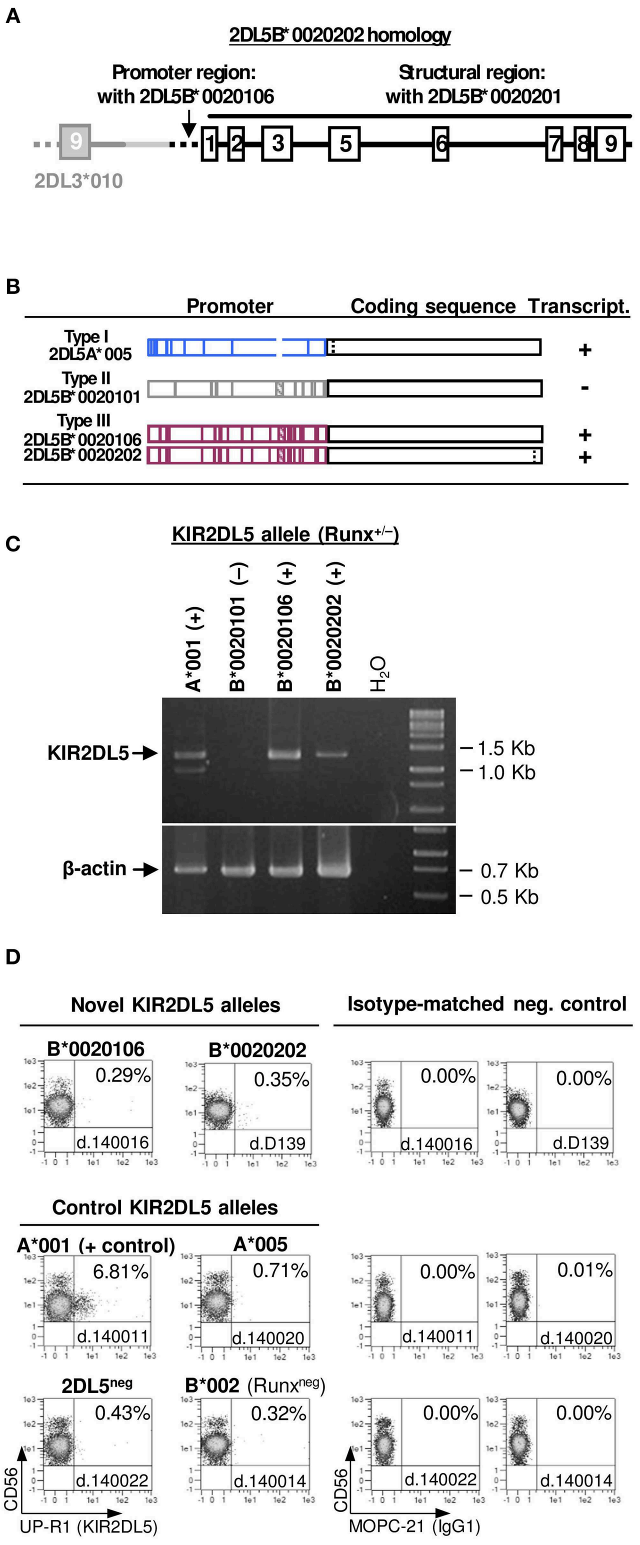


**FIGURE 7 |** Characterization of *KIR2DL5B*0020106* and **0020202*. **(A)** Gene structure of *KIR2DL5B*0020202*, including homology to other *KIR2DL5* alleles. **(B)** Comparison of *KIR2DL5* alleles carrying nearly identical coding sequences but highly divergent promoter regions; vertical lines represent polymorphisms distinguishing those alleles. Gene transcription or silencing is indicated on the right side. **(C)** RT-PCR assay showing *KIR2DL5B*0020106* and **0020202* transcription, in contrast with their common, silent homologue *KIR2DL5B*0020101*. Presence or absence of an intact RUNX binding site in the proximal promoter is indicated for each allele. **(D)** A KIR2DL5 product is undetectable on the surface of NK cells transcribing *KIR2DL5B*0020106* and **0020202*, as we reported previously for *KIR2DL5A*005*, which encodes an identical mature polypeptide.

## DISCUSSION

In-depth genomic studies have established the variable organization of the human *KIR* complex, defining gene motifs and extended haplotypes fixed in our species, and a series of variations from those, mostly generated by asymmetric (non-allelic) homologous recombination (4, 6–10, 26, 28, 29). In addition, sequence analyses based on Sanger and, more recently, second generation methods, have revealed the diversity of alleles commonly found in each of those *KIR* haplotypes (23, 24, 30–32, 34–39).

In parallel, the *KIR*-gene profiles of multiple populations worldwide have been explored in the last two decades, revealing that, whilst many *KIR* haplotypes and alleles are shared by humans of all ethnicities, notorious differences in their distribution exist, unveiling the evolutionary connections between human groups, and the variable selective pressures exerted on them by the environment (16, 40). However, it is noteworthy that many published population studies have not benefited sufficiently from knowledge on *KIR* polymorphism gained from genomic studies, often reporting only rough analyses of gene content and basic classifications on the "B-ness" or "A-ness" of *KIR* haplotypes.

We have tried to contribute to fill that gap by studying a sample of Spanish individuals by means of: (i) A combination of rapid and advanced methods for *KIR*-gene profiling that inform, not only of presence/absence of *KIR* genes, but also of isoforms associated with defined haplotypes, and recombinants that mark contracted/expanded haplotypes; (ii) Parsimonious interpretation of *KIR*-gene profiles based on accumulated knowledge on common and well-defined haplotypes and alleles; and (iii) Basic molecular characterization of a minority of *KIR*-gene profiles not fitting with known arrangements. Limitations of our study are that polymorphism has, in general, not been defined at an allelic level; and that our parsimonious approach might theoretically overlook part of the existing diversity. To mitigate the latter limitation we have screened certain specific recombinations in donors with compatible profiles, thus identifying individuals in whom variant structures were concealed by the accompanying genes.

By applying systematically this approach, we have managed to explain all 38 gene profiles found in 414 individuals, inferring their haplotype structures and incorporating them to analysis of *KIR*-gene distribution. We have thus determined the detailed distribution of the three fixed centromeric and three telomeric motifs; of six expanded or contracted *KIR*-gene arrangements characterized previously by us and others, seen in 22 individuals (5.31%); and two novel haplotypes never detected before in Caucasoids. These novel arrangements are associated with new *KIR* alleles, and they show combined features of B- and A-haplotypes.

Transcribed *KIR2DL5B**002* alleles, contrasting with common silent ones, were found within a *2DL5B-2DS3* cluster inserted in a centromeric *2DL3-2DP1-2DL1* motif, thus converting it in a cen-B haplotype. Such structure had been previously reported in individuals of African origin (8, 9, 32). Transcription of *KIR2DL5B**0020106* and **0020202* provides further confirmation to our hypothesis that an intact RUNX binding site in the proximal promoter is essential for *KIR*-gene expression, whilst its mutation determines epigenetic silencing. The biological significance of transcribed *KIR2DL5B**002* alleles is, however, uncertain since, like KIR2DL5A*005 (**Figure 7**), the encoded receptor appears not to reach the cell surface and is possibly retained intracellularly (56). The fact that nearly identical coding sequences are preceded by three highly divergent promoter sequences in alleles of two paralog genes (*2DL5A**005*, *2DL5B**0020101* and *2DL5B**0020106/0020202*) illustrates the role of recombination in *KIR*-gene evolution.

The second novel allele, *KIR2DL3**033*, and its associated haplotype do not fit comfortably with the current classification and designation of *KIR* genes and haplotypes, posing a puzzling nomenclature issue. Its long inhibitory tail, homologous to that of common *KIR2DL3* alleles, should warrant its designation as a *KIR2DL*. Such name, however, challenges an unwritten rule of assigning hybrid *KIR* to the locus contributing the extracellular portion (e.g. *2DS2**005*, a *2DS2/2DS3* hybrid, or *3DL1**060*, a *3DL1/3DL2* chimera), and it obviates that most of the gene (∼11 of 14 Kbp) and of the encoded molecule (all the ectodomain) are actually identical to those of *KIR2DS2*, and will be detected as such by most current genotyping and phenotyping assays. This may result in ambiguous or conflicting profiles, which can be sorted out by assays that, like the one we have used (**Figure 6A**), target specifically the *KIR2DL3**033* recombination spot in intron 6. Furthermore, whereas recombinations of B- and A-haplotypes normally yield B-haplotypes, gene content of the hybrid haplotype bearing *2DL3**033* adjusts to the definition of A-haplotypes, its "B-ness" being perceived only when the actual sequence of its centromeric genes (*KIR 3DL3* and *2DL3*) is considered. On the other hand, assigning the new structure to the *KIR2DS2* gene would better reflect its origin, but would imply designating a long-tailed KIR with an "S" symbol aimed at distinguishing short-tailed, activating, KIR. To circumvent such and similar inconsistencies, the official KIR nomenclature might consider in the future the use of dedicated names to describe the hybrid nature of recombinant KIR genes (e.g., KIR2DS2/L3, KIR3DL1/L2, et cetera).

In summary, we consider that our study provides a representative and precise estimate of the KIR structures seen in the Spanish population, and extends our understanding of their complexity in South European Caucasoids. We expect that our results, and the approach we followed to obtain them, will enable better-founded studies of KIR in populations, and in health conditions in which their genetic diversity is deemed relevant.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the EMBL/GenBank/DDBJ HG931348, LT604077.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comité Ético de Investigación con Medicamentos, Hospital Universitario Puerta de Hierro, Majadahonda. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

EC designed and performed experiments, analyzed and interpreted data, and wrote the manuscript. MM performed experiments, analyzed and interpreted data, and revised the manuscript. NG-L, AM, and ML-B contributed samples, and revised the manuscript. CV designed the study, directed research and wrote the manuscript.

## FUNDING

This work was supported by grants SAF2016-80363-C2-1/2-R (Spanish Ministry of Economy and Competitiveness (MINECO)-FEDER, EU), and PCIN-2015-191-C02-02 (FP7-MINECO Infect-ERA program) to CV and ML-B. MM was supported by *Fundación Asociación Española Contra el Cáncer* (AECC) grant GCB15152947MELE. EC was supported, successively, by grants SAF2010–22153-C03, *Fundació La Marató de TV3* 121531, and PCIN-2015-191-C02-02. AM was supported by Instituto de Salud Carlos III (ISCIII)/FEDER grant PI19/00328. CV was supported in 2019 and 2020 by the nominative grant from *Consejería de Sanidad de la Comunidad de Madrid* to *Fundación de Investigación Biomédica Puerta de Hierro*.

## ACKNOWLEDGMENTS

We thank all individuals who generously donated samples for this study. We also thank Dr. Rosario de Pablo, Rosario Solís, and Esperanza Rodríguez (former members of the HLA lab, Inmunología, Hospital Universitario Puerta de Hierro), Gemma Heredia (University Pompeu Fabra), and members of the central facilities of *Instituto de Investigación Sanitaria Puerta de Hierro Segovia de Arana*, for expert support.

## REFERENCES

1. Vilches C, Parham P. KIR: diverse, rapidly evolving receptors of innate and adaptive immunity. *Annu Rev Immunol.* (2002) 20:217–51. doi: 10.1146/annurev.immunol.20.092501.134942
2. Falco M, Pende D, Munari E, Vacca P, Mingari MC, Moretta L. Natural killer cells: from surface receptors to the cure of high-risk leukemia (Ceppellini Lecture). *HLA.* (2019) 93:185–94. doi: 10.1111/tan.13509
3. Wroblewski EE, Parham P, Guethlein LA. Two to tango: co-evolution of hominid natural killer cell receptors and MHC. *Front Immunol.* (2019) 10:177. doi: 10.3389/fimmu.2019.00177
4. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci USA.* (2000) 97:4778–4783. doi: 10.1073/pnas.080588597
5. Sambrook JG, Bashirova A, Andersen H, Piatak M, Vernikos GS, Coggill P, et al. Identification of the ancestral killer immunoglobulin-like receptor gene in primates. *BMC Genomics.* (2006) 7:209. doi: 10.1186/1471-2164-7-209
6. Martin AM, Freitas EM, Witt CS, Christiansen FT. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics.* (2000) 51:268–80. doi: 10.1007/s002510050620
7. Hsu KC, Chida S, Geraghty DE, Dupont B. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism. *Immunol Rev.* (2002) 190:40–52. doi: 10.1034/j.1600-065X.2002.19004.x
8. Pyo CW, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different patterns of evolution in the centromeric and telomeric regions of group A and B haplotypes of the human killer cell Ig-like receptor locus. *PLoS ONE.* (2010) 5:e15115. doi: 10.1371/journal.pone.0015115
9. Pyo CW, Wang R, Vu Q, Cereb N, Yang SY, Duh FM, et al. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics.* (2013) 14:89. doi: 10.1186/1471-2164-14-89
10. Roe D, Vierra-Green C, Pyo CW, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun.* (2017) 18:127–34. doi: 10.1038/gene.2017.10
11. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, et al. Human diversity in killer cell inhibitory receptor genes. *Immunity.* (1997) 7:753–63. doi: 10.1016/S1074-7613(00)80394-5
12. Vilches C, Rajalingam R, Uhrberg M, Gardiner CM, Young NT, and Parham P. KIR2DL5, a novel killer-cell receptor with a D0-D2 configuration of Ig-like domains. *J Immunol.* (2000) 164:5797–804. doi: 10.4049/jimmunol.164.11.5797
13. Marsh SG, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Tissue Antigens.* (2003) 62:79–86. doi: 10.1034/j.1399-0039.2003.00072.x
14. Maxwell LD, Wallace A, Middleton D, Curran MD. A common KIR2DS4 deletion variant in the human that predicts a soluble KIR molecule analogous to the KIR1D molecule observed in the rhesus monkey. *Tissue Antigens.* (2002) 60:254–8. doi: 10.1034/j.1399-0039.2002.600307.x
15. Cisneros E, Moraru M, Gomez-Lozano N, Lopez-Botet M, Vilches C. KIR2DL5: an orphan inhibitory receptor displaying complex patterns of polymorphism and expression. *Front Immunol.* (2012) 3:289. doi: 10.3389/fimmu.2012.00289
16. Parham P, Guethlein LA. Genetics of natural killer cells in human health, disease, and survival. *Annu Rev Immunol.* (2018) 36:519–48. doi: 10.1146/annurev-immunol-042617-053149
17. Ordóñez D, Meenagh A, Gómez-Lozano N, Castaño J, Middleton D, Vilches C. Duplication, mutation and recombination of the human orphan gene KIR2DS3 contribute to the diversity of KIR haplotypes. *Genes Immun.* (2008) 9:431–7. doi: 10.1038/gene.2008.34
18. Gómez-Lozano N, Gardiner CM, Parham P, Vilches C. Some human KIR haplotypes contain two KIR2DL5 genes: KIR2DL5A and KIR2DL5B. *Immunogenetics.* (2002) 54:314–9. doi: 10.1007/s00251-002-0476-2
19. Norman PJ, Carrington CV, Byng M, Maxwell LD, Curran MD, Stephens HA, et al. Natural killer cell immunoglobulin-like receptor (KIR) locus profiles in African and South Asian populations. *Genes Immun.* 3:86–95. doi: 10.1038/sj.gene.6363836
20. Gomez-Lozano N, de Pablo R, Puente S, and Vilches C. (2003). Recognition of HLA-G by the NK cell receptor KIR2DL4 is not essential for human reproduction. *Eur J Immunol.* 33:639–44. doi: 10.1002/eji.200323741
21. Martin MP, Bashirova A, Traherne J, Trowsdale J, Carrington M. Cutting edge: expansion of the KIR locus by unequal crossing over. *J Immunol.* (2003) 171:2192–5. doi: 10.4049/jimmunol.171.5.2192
22. Gomez-Lozano N, Estefania E, Williams F, Halfpenny I, Middleton D, Solis R, et al. The silent KIR3DP1 gene (CD158c) is transcribed and might encode a secreted receptor in a minority of humans, in whom the KIR3DP1, KIR2DL4 and KIR3DL1/KIR3DS1 genes are duplicated. *Eur J Immunol.* (2005) 35:16–24. doi: 10.1002/eji.200425493
23. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet.* (2010) 19:737–51. doi: 10.1093/hmg/ddp538
24. Hou L, Chen M, Ng J, Hurley CK. Conserved KIR allele-level haplotypes are altered by microvariation in individuals with European ancestry. *Genes Immun.* (2012) 13:47–58. doi: 10.1038/gene.2011.52
25. Valiante NM, Uhrberg M, Shilling HG, Lienert-Weidenbach K, Arnett KL, D'Andrea A, et al. Functionally and structurally distinct NK cell receptor repertoires in the peripheral blood of two human donors. *Immunity.* (1997) 7:739–51. doi: 10.1016/S1074-7613(00)80393-3
26. Vilches C, Gardiner CM, Parham P. Gene structure and promoter variation of expressed and non-expressed variants of the KIR2DL5 gene. *J Immunol.* (2000) 165:6416–21. doi: 10.4049/jimmunol.165.11.6416
27. Martin MP, Single RM, Wilson MJ, Trowsdale J, Carrington M. KIR haplotypes defined by segregation analysis in 59 Centre d'Etude Polymorphisme Humain (CEPH) families. *Immunogenetics.* (2008) 60:767–74. doi: 10.1007/s00251-008-0334-y
28. Selvaraj S, Schmitt AD, Dixon JR, Ren B. Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq. *BMC Genomics.* (2015) 16:900. doi: 10.1186/s12864-015-1949-7
29. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. *Am J Hum Genet.* (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023
30. Meenagh A, Williams F, Sleator C, Halfpenny IA, Middleton D. Investigation of killer cell immunoglobulin-like receptor gene diversity V. KIR3DL2. *Tissue Antigens.* (2004) 64:226–34. doi: 10.1111/j.1399-0039.2004.00272.x
31. Middleton D, Meenagh A, Gourraud PA. KIR haplotype content at the allele level in 77 Northern Irish families. *Immunogenetics.* (2007) 59:145–58. doi: 10.1007/s00251-006-0181-7
32. Hou L, Chen M, Jiang B, Wu D, Ng J, Hurley CK. Thirty allele-level haplotypes centered around KIR2DL5 define the diversity in an African American population. *Immunogenetics.* (2010) 62:491–8. doi: 10.1007/s00251-010-0458-8
33. Hollenbach JA, Nocedal I, Ladner MB, Single RM, Trachtenberg EA. Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations. *Immunogenetics.* (2012) 64:719–37. doi: 10.1007/s00251-012-0629-x
34. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Res.* (2012) 22:1845–54. doi: 10.1101/gr.137976.112
35. Vierra-Green C, Roe D, Hou L, Hurley CK, Rajalingam R, Reed E, et al. Allele-level haplotype frequencies and pairwise linkage disequilibrium for 14 KIR loci in 506 European-American individuals. *PLoS ONE.* (2012) 7:e47491. doi: 10.1371/journal.pone.0047491
36. Pontikos N, Smyth DJ, Schuilenburg H, Howson JM, Walker NM, Burren OS, et al. A hybrid qPCR/SNP array approach allows cost efficient assessment of KIR gene copy numbers in large samples. *BMC Genomics.* (2014) 15:274. doi: 10.1186/1471-2164-15-274
37. Roberts CH, Jiang W, Jayaraman J, Trowsdale J, Holland MJ, Traherne JA. Killer-cell Immunoglobulin-like Receptor gene linkage and copy number variation analysis by droplet digital PCR. *Genome Med.* (2014) 6:20. doi: 10.1186/gm537
38. Jiang W, Johnson C, Simecek N, Lopez-Alvarez MR, Di D, Trowsdale J, et al. qKAT: a high-throughput qPCR method for KIR gene copy

number and haplotype determination. *Genome Med.* (2016) 8:99. doi: 10.1186/s13073-016-0358-0

39. Wagner I, Schefzyk D, Pruschke J, Schofl G, Schone B, Gruber N, et al. Allele-level KIR genotyping of more than a million samples: workflow, algorithm, and observations. *Front Immunol.* (2018) 9:2843. doi: 10.3389/fimmu.2018.02843
40. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, da Silva AL, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* (2015). 43:D784–8. doi: 10.1093/nar/gku1166
41. Gómez-Lozano N, Trompeter HI, de Pablo R, Estefanía E, Uhrberg M, Vilches C. Epigenetic silencing of potentially functional KIR2DL5 alleles: implications for the acquisition of KIR repertoires by NK cells. *Eur J Immunol.* (2007) 37:1954–65. doi: 10.1002/eji.200737277
42. Vilches C, Castano J, Gomez-Lozano N, Estefania E. Facilitation of KIR genotyping by a PCR-SSP method that amplifies short DNA fragments. *Tissue Antigens.* (2007) 70:415–22. doi: 10.1111/j.1399-0039.2007.00923.x
43. Ordóñez D, Gómez-Lozano N, Rosales L, Vilches C. Molecular characterisation of KIR2DS2*005, a fusion gene associated with a shortened KIR haplotype. *Genes Immunity.* (2011) 12:544–51. doi: 10.1038/gene.2011.35
44. Artavanis-Tsakonas K, Eleme K, McQueen KL, Cheng NW, Parham P, Davis DM, et al. Activation of a subset of human NK cells upon contact with *Plasmodium falciparum*-infected erythrocytes. *J Immunol.* (2003) 171:5396–405. doi: 10.4049/jimmunol.171.10.5396
45. Norman PJ, Abi-Rached L, Gendzekhadze K, Hammond JA, Moesta AK, Sharma D, et al. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. *Genome Res.* (2009). 19:757–69. doi: 10.1101/gr.085738.108
46. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* (2001) 68:978–89. doi: 10.1086/319501
47. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* (2003) 73:1162–9. doi: 10.1086/379378
48. Gaunt TR, Rodriguez S, Day IN. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool "CubeX." *BMC Bioinformatics.* (2007) 8:428. doi: 10.1186/1471-2105-8-428
49. Estefanía E, Flores R, Gómez-Lozano N, Aguilar H, López-Botet M, Vilches C. Human KIR2DL5 is an inhibitory receptor expressed on the surface of NK and T lymphocyte subsets. *J Immunol.* (2007) 178:4402–10. doi: 10.4049/jimmunol.178.7.4402
50. Leaton LA, Shortt J, Kichula KM, Tao S, Nemat-Gorgani N, Mentzer AJ, et al. Conservation, extensive heterozygosity, and convergence of signaling potential all indicate a critical role for KIR3DL3 in higher primates. *Front Immunol.* (2019) 10:24. doi: 10.3389/fimmu.2019.00024
51. Diaz-Pena R, Vidal-Castineira JR, Moro-Garcia MA, Alonso-Arias R, Castro-Santos P. Significant association of the KIR2DL3/HLA-C1 genotype with susceptibility to Crohn's disease. *Hum Immunol.* (2016) 77:104–9. doi: 10.1016/j.humimm.2015.10.020
52. Castano-Nunez A, Montes-Cano MA, Garcia-Lozano JR, Ortego-Centeno N, Garcia-Hernandez FJ, Espinosa G, et al. Association of functional polymorphisms of KIR3DL1/DS1 with Behcet's disease. *Front Immunol.* (2019) 10:2755. doi: 10.3389/fimmu.2019.02755
53. Closa L, Vidal F, Herrero MJ, Caro JL. Distribution of human killer cell immunoglobulin-like receptors and ligands among blood donors of Catalonia. *HLA*. (2019) 95:179–88. doi: 10.1111/tan.13754
54. Bono M, Pende D, Bertaina A, Moretta A, Della Chiesa M, Sivori S, et al. Analysis of KIR3DP1 polymorphism provides relevant information on centromeric KIR gene content. *J Immunol.* (2018) 201:1460–7. doi: 10.4049/jimmunol.1800564
55. Nakimuli A, Chazara O, Farrell L, Hiby SE, Tukwasibwe S, Knee O, et al. Killer cell immunoglobulin-like receptor (KIR) genes and their HLA-C ligands in a Ugandan population. *Immunogenetics.* (2013) 65:765–75. doi: 10.1007/s00251-013-0724-7
56. Cisneros E, Estefania E, Vilches C. Allelic polymorphism determines surface expression or intracellular retention of the human NK cell receptor KIR2DL5A (CD158f). *Front Immunol.* (2017) 7:698. doi: 10.3389/fimmu.2016.00698

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

ORIGINAL RESEARCH
published: 02 April 2020
doi: 10.3389/fimmu.2020.00556

# KIR Variation in Iranians Combines High Haplotype and Allotype Diversity With an Abundance of Functional Inhibitory Receptors

Claudia Alicata[1†], Elham Ashouri[2,3,4,5†], Neda Nemat-Gorgani[2,3,6], Lisbeth A. Guethlein[2,3], Wesley M. Marin[7], Sudan Tao[8,9], Lorenzo Moretta[1], Jill A. Hollenbach[7], John Trowsdale[6], James A. Traherne[6], Abbas Ghaderi[5], Peter Parham[2,3] and Paul J. Norman[2,3,9*]

[1] Department of Immunology, IRCCS Bambino Gesù Children's Hospital, Rome, Italy, [2] Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, United States, [3] Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, United States, [4] Hematology-Oncology and Stem Cell Transplantation Research Center, Tehran University of Medical Sciences, Tehran, Iran, [5] School of Medicine, Shiraz Institute for Cancer Research, Shiraz University of Medical Sciences, Shiraz, Iran, [6] Division of Immunology, Department of Pathology, University of Cambridge, Cambridge, United Kingdom, [7] Department of Neurology, University of California, San Francisco, San Francisco, CA, United States, [8] Blood Center of Zhejiang Province, Hangzhou, China, [9] Division of Personalized Medicine, Department of Immunology and Microbiology, University of Colorado, Anschutz Medical Campus, Aurora, CO, United States

OPEN ACCESS

**Edited by:**
Martin Maiers,
National Marrow Donor Program,
United States

**Reviewed by:**
Ketevan Gendzekhadze,
City of Hope National Medical Center,
United States
Johannes Schetelig,
Universitätsklinikum Carl Gustav
Carus, Germany

***Correspondence:**
Paul J. Norman
paul.norman@ucdenver.edu

[†] These authors have contributed equally to this work

**Specialty section:**
This article was submitted to
NK and Innate Lymphoid Cell Biology,
a section of the journal
Frontiers in Immunology

**Received:** 04 December 2019
**Accepted:** 11 March 2020
**Published:** 02 April 2020

**Citation:**
Alicata C, Ashouri E,
Nemat-Gorgani N, Guethlein LA,
Marin WM, Tao S, Moretta L,
Hollenbach JA, Trowsdale J,
Traherne JA, Ghaderi A, Parham P
and Norman PJ (2020) KIR Variation in
Iranians Combines High Haplotype
and Allotype Diversity With an
Abundance of Functional Inhibitory
Receptors. Front. Immunol. 11:556.
doi: 10.3389/fimmu.2020.00556

Natural killer (NK) cells are innate lymphocytes that eliminate infected and transformed cells. They discriminate healthy from diseased tissue through killer cell Ig-like receptor (KIR) recognition of HLA class I ligands. Directly impacting NK cell function, *KIR* polymorphism associates with infection control and multiple autoimmune and pregnancy syndromes. Here we analyze *KIR* diversity of 241 individuals from five groups of Iranians. These five populations represent Baloch, Kurd, and Lur, together comprising 15% of the ethnically diverse Iranian population. We identified 159 *KIR* alleles, including 11 not previously characterized. We also identified 170 centromeric and 94 telomeric haplotypes, and 15 different *KIR* haplotypes carrying either a deletion or duplication encompassing one or more complete *KIR* genes. As expected, comparing our data with those representing major worldwide populations revealed the greatest similarity between Iranians and Europeans. Despite this similarity we observed higher frequencies of *KIR3DL1*001* in Iran than any other population, and the highest frequency of HLA-B*51, a Bw4-containing allotype that acts as a strong educator of *KIR3DL1*001*$^+$ NK cells. Compared to Europeans, the Iranians we studied also have a reduced frequency of *3DL1*004*, which encodes an allotype that is not expressed at the NK cell surface. Concurrent with the resulting high frequency of strong viable interactions between inhibitory KIR and polymorphic HLA class I, the majority of *KIR-A* haplotypes characterized do not express a functional activating receptor. By contrast, the most frequent *KIR-B* haplotype in Iran expresses only one functional inhibitory KIR and the maximum number of activating KIR. This first complete, high-resolution, characterization of the *KIR* locus of Iranians will form a valuable reference for future clinical and population studies.

**Keywords: NK cells, KIR, HLA class I, Iranian populations, immune diversity**

## INTRODUCTION

Natural killer (NK) cells are essential for human immunity to infection and cancer, and for successful reproduction (1, 2). To discriminate diseased from healthy tissue cells, NK cells express an array of inhibiting and activating cell surface receptors (3, 4). Prominent among these receptors are the killer cell immunoglobulin like receptors (KIR), which educate and modulate NK cell function through interaction with HLA class I (5, 6). KIR are highly polymorphic, a genetically-determined variation that directly impacts NK cell function, and susceptibility to disease (7, 8).

The *KIR* locus spans 150–350 kbp of chromosome 19q13.4 (9). The locus is distinguished by structural and sequence diversity of the 13 constituent genes (*KIR2DL1*, *KIR2DL2/L3*, *KIR2DL4*, *KIR2DL5A*, *KIR2DL5B*, *KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5*, *KIR3DL1/S1*, *KIR3DL2*, and *KIR3DL3*) and two pseudogenes (*KIR2DP1* and *KIR3DP1*) (10, 11). KIR have either two (2D) or three (3D) specificity-determining immunoglobulin-like domains and a long (L) or short (S) tail (12). KIR having a long cytoplasmic tail are inhibitory, whereas those having a short cytoplasmic tail are activating. The one exception is KIR2DL4, which can exhibit inhibitory or activating function (13–15). The *KIR* locus segregates in two main haplotype forms that are maintained in all human populations by balancing selection (16). The *KIR-A* haplotypes have a fixed number of predominantly inhibitory receptors, whereas the *KIR-B* haplotypes are characterized by a variable number of both activating and inhibitory receptors. Further distinguishing *KIR-A* and *-B* haplotypes are their characteristic alleles (17). *KIR-A* haplotypes are associated with controlling infectious disease and cancer, but confer susceptibility to reproductive disorders, whereas *KIR-B* haplotypes are associated with protection from reproductive disorders (1, 7). Additionally, haploidentical transplantation therapy for leukemia has an increased success rate when the stem cell donors carry *KIR-B* haplotypes (18, 19).

Polymorphism of *KIR* affects cell surface expression, ligand specificity, ligand binding strength, and intracellular signaling (20–25). All these factors affect the capacity of NK cells to recognize and kill target cells. Because both *KIR* polymorphism and associated diseases are unevenly distributed worldwide, it is critical to fully gauge the genetic diversity of *KIR* in well-defined human populations. Despite this importance, only a few populations have been studied to high resolution, principally due to the complexity of the *KIR* locus. These studies have focused on representative populations of Amerindians (Yucpa) (26), divergent African groups (27–29), Europeans (30, 31), East Asians (Japanese) (32), and Oceanians (Māori) (33). Together, they show how *KIR* allele and haplotype diversity varies dramatically between human populations and highlight the importance of extending *KIR* allele analysis to represent all ethnicities and geographical areas. In this regard, our recent analysis of *HLA* allelic diversity in Iran revealed the highest frequencies of *HLA-B*51:01* worldwide (34). HLA-B*51 contains the Bw4 epitope and interacts KIR3DL1 (35, 36). In the present study we fully characterize *KIR* locus diversity of the same cohort of Iranian individuals.

## MATERIALS AND METHODS

### Study Population

The *KIR* locus diversity of three indigenous Iranian populations was determined to high-resolution by analyzing genomic DNA from 241 healthy unrelated donors (34). The populations studied represent Baloch, Kurd, and Lur, together comprising 15% (12 million individuals) of the ethnically diverse Iranian population. Studied were 160 Lurs and 48 Kurds, from the Zagros Mountains at the west of Iran, and 33 Baloch from the southeast of Iran. The Lur population included 64 individuals from the city of Khoramabad in the province of Lorestan, 81 from Yasuj in the province of Kohgiluyeh and Boyer-Ahmad, and 15 from Lordegan in the province of Chaharmahal and Bakhtiari. The samples from Kurds were collected from the city of Sanandaj. With the exception of the Baloch, the *HLA class I* alleles were described previously (34). The ancestors of every individual studied had been part of their respective population for at least two generations. Sample collection was approved by the Medical Research Ethics Committee of Shiraz University of Medical Sciences. All participants gave informed consent. Banked, de-identified samples were used for this study.

### Library Preparation and Enrichment

Genomic DNA was prepared by shearing with sonication and the *KIR* genomic region was enriched from the genomic libraries using a pool of oligonucleotide probes as described (37). The enriched fragments were subjected to paired-end sequencing using Illumina's MiSeq instrument and V3 sequencing chemistry (Illumina, La Jolla CA). The sequencing read length was 2 × 300 bp.

### Next Generation Sequence Data Processing and Analysis

Sequence reads specific to the *KIR* region were identified and harvested using Bowtie 2 (38). KIR genotyping was performed using the Pushing Immunogenetics to the Next Generation pipeline (37). This pipeline generates a high-resolution *KIR* gene content and allele level genotype. It can also identify previously unreported single nucleotide polymorphisms (SNPs) and recombinant alleles. Novel allele sequences were analyzed by visual inspection: reads specific to the relevant gene were isolated by bioinformatics filtering, aligned to the closest reference allele using MIRA 4.0.2 (39), and inspected using Gap4 of the Staden package (40) or Integrative Genomics Viewer (41).

### Allele and Haplotype Frequencies

Allele frequencies were calculated by direct counting. The composition and frequencies of *KIR* haplotypes were determined at the allelic level using PHASE 2.1 (42). The following parameters for PHASE 2.1 were used: –f1, –x5, and –d1. Because of the high rate of recombination between *KIR3DP1* and *KIR2DL4* (9), we performed two separate PHASE runs, one for the *KIR* genes of the centromeric region and one for telomeric *KIR* genes. Genes analyzed were *KIR3DL3, 2DS2, 2DL2/3, 2DL5A* and *B, 2DS3/5, 2DP1, 2DL1, 2DL4, 3DL1/S1,*

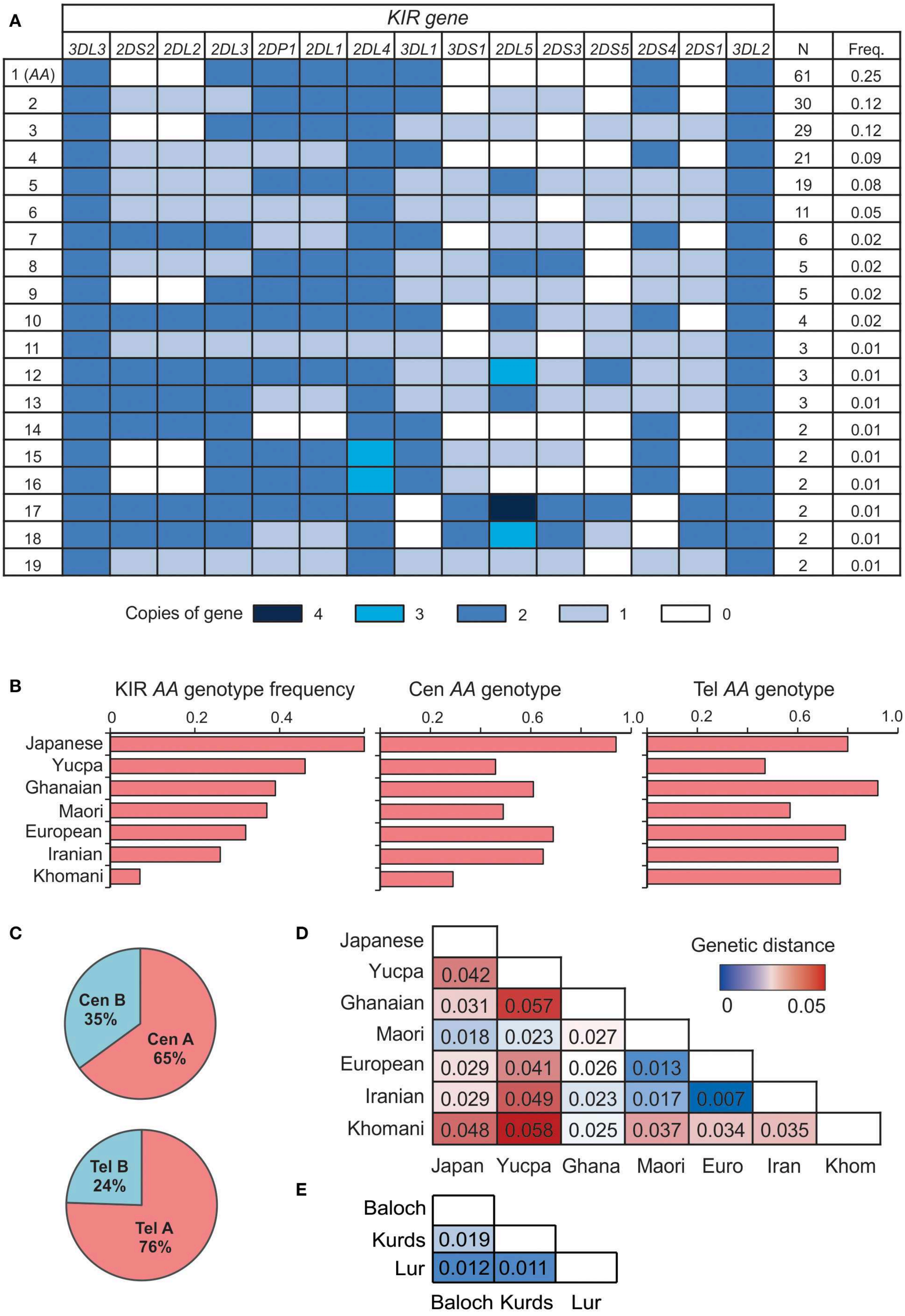


**FIGURE 1 |** Iranian *KIR* genotypes resemble those of Europeans. **(A)** Shows *KIR* gene copy-number genotypes ordered by the total number observed across the five Iranian populations. Only genotypes present in more than one individual are shown. Colored boxes indicate the number of copies of the gene, as given in the key

*(Continued)*

**FIGURE 1 |** below. A white box indicates the gene is absent. **(B)** Frequency of the full *KIR AA* (left), *Cen AA* (center), and *Tel AA* (right) genotypes across seven representative world populations, ordered by frequency of full *KIR AA*. The populations are Japanese (*N* = 115), Yucpa Amerindians (*N* = 61), Ghanaians from West Africa (*N* = 131), Maori from New Zealand (*N* = 49), Europeans (*N* = 378), Iranians (*N* = 241), and Khomani from South Africa (*N* = 79). **(C)** Frequency of centromeric (Cen) and telomeric (Tel) *KIR A* and *B* haplotypes in the five Iranian populations combined. **(D)** Genetic distance between the combined Iranian population and the six representative, populations from **(B)**. *KIR* genes included are those genotyped in all the populations shown *2DL1-4, 2DS1-5, 3DL1/S1, 3DL2*. **(E)** Genetic distance between the three Iranian populations.

*2DS1, 2DS4*, and *3DL2*. For each haplotype we calculated the frequency by direct counting.

## Statistical Analysis

Hardy-Weinberg equilibrium proportions was examined using Fisher's exact test. Differences in frequency amongst populations were tested using $\chi^2$ with Bonferroni correction for the number of alleles at the respective locus. Fisher's and $\chi^2$ test as well as Mann-Whitney *U*-test were implemented using GraphPad Prism 7.05.

## Genetic Distance

The genetic distance between Iranians and other populations was calculated using the Cavalli-Sforza model (43), implemented in GENETIX 4.05 (https://kimura.univ-montp2.fr/genetix/). The populations used were Japanese (*N* = 115) (32), Yucpa Amerindians (*N* = 61) (26), Ghanaians from West Africa (*N* = 131) (27), Māori from New Zealand (*N* = 49) (33), Europeans (*N* = 378) (44), and Khomani from South Africa (*N* = 79) (28).

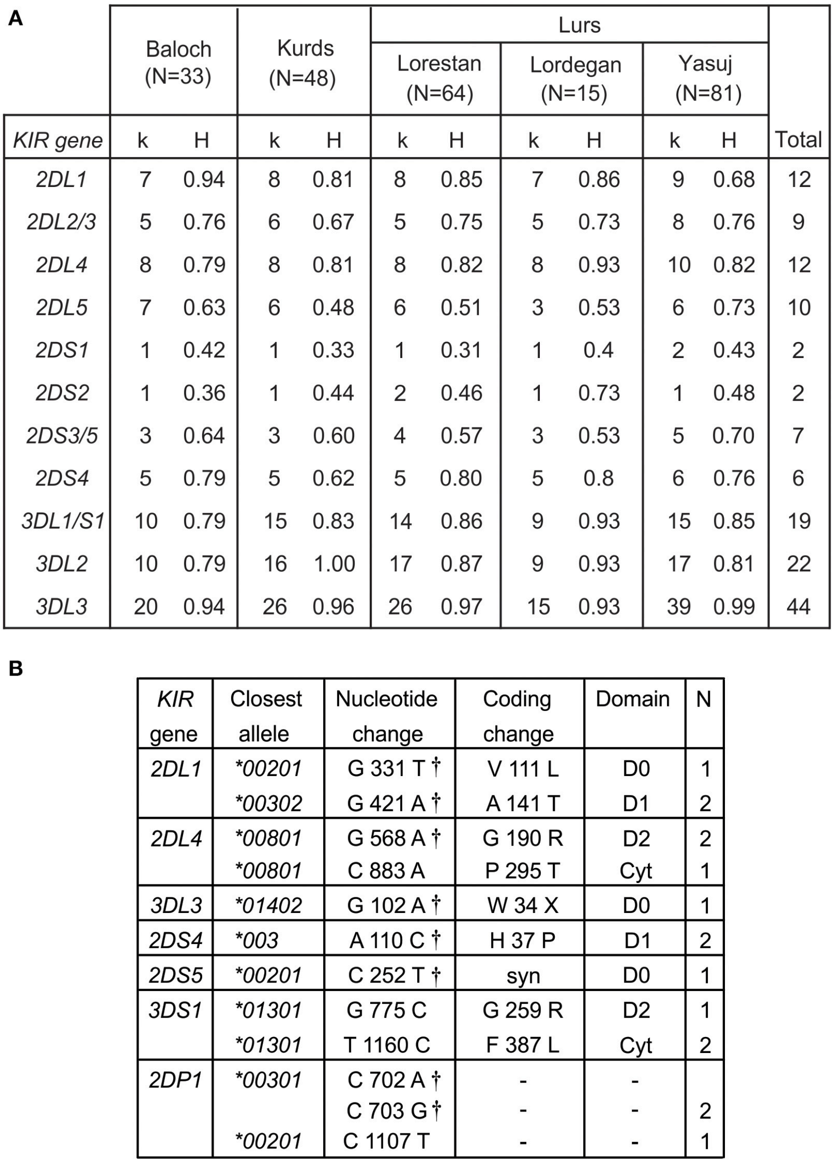
# RESULTS

We sequenced the *KIR* genes of 241 individuals from five populations, representing three groups of Iranians; the Lurs, Kurds and Baloch. The Lurs comprised individuals from the cities of Khoramabad, Lordegan and Yasuj. Only two *KIR* genes were present as two copies (2N) in every individual, *KIR3DL3* at the centromeric end of the *KIR* locus and *KIR3DL2* at the telomeric end (**Figure 1A**). Every *KIR* haplotype in this Iranian cohort is therefore flanked by these two framework genes. The third framework gene is *KIR2DL4* (9). Because eight individuals have only one copy of *KIR2DL4* and six have three copies, *KIR2DL4* is not present on every Iranian *KIR* haplotype (**Figure 1A**), but is likely duplicated on some haplotypes and deleted from others (45). The frequency of *KIR-A* haplotypes varies across populations (16, 46). In the combined study population, the frequency of *KIR AA* genotypes observed is 25% (**Figure 1B**). C*entromeric KIR-A* haplotypes are present at 65% and *telomeric KIR-A* haplotypes at 76% (**Figure 1C**). The frequency of *KIR A* haplotype homozygotes closely matches that of European populations (**Figure 1B**), as do genetic distance measurements calculated from the *KIR* genotype data (**Figure 1D**). In conclusion, the number and distribution of *KIR* gene content haplotypes in Iranians closely resemble those present in Europeans.

We determined the *KIR* allele frequencies in the five Iranian populations. These data are shown in **Supplementary Material A** and summarized in **Figure 2**. The allele frequencies were consistent with Hardy-Weinberg equilibrium. Among the five populations, we identified 115 inhibitory KIR alleles and 18 activating KIR alleles (**Figure 2A**). Also present are 12 *KIR2DL4* alleles. Inhibitory KIR are highly polymorphic in Iranians, with 12 *KIR2DL1*, 9 *KIR2DL2/L3*, 10 *KIR2DL5*, 22 *KIR3DL2*, 18 *KIR3DL1*, and 44 *KIR3DL3* alleles observed in total. As in other populations, the activating KIR are less polymorphic than the inhibitory KIR, with 7 *KIR2DS4* alleles, 8 *KIR2DS3/5* alleles, two *KIR2DS1* and two *KIR2DS2* alleles and one *KIR3DS1* allele being seen (**Figure 2A**). In Iranians, *KIR3DL3* is the most polymorphic *KIR* gene, whereas *KIR2DS1* and *KIR2DS2* are the least variable. In the combined population analyzed, the most frequent allele for each of the inhibitory KIR are *2DL1*00302*, *2DL2*00101*, *2DL3*00101*, *3DL1*00101*, *3DL2*00101*, and *3DL3*00301,* respectively (**Supplementary Material A**). The most common KIR2DL4 and KIR2DS4 alleles are *2DL4*00801* and *2DS4*003*, respectively. For the other activating KIR (KIR2DS1-3 and 2DS5) as well as KIR2DL5, the most frequent allele observed is absence of the gene (**Supplementary Material A**). We did not observe any statistically significant difference in frequency of any specific *KIR* allele between Kurs and Lurs. We identified 11 novel *KIR* alleles, eight being defined by amino acid substitutions, one by a synonymous substitution and two by substitutions in *KIR2DP1* pseudogene (**Figure 2B**). All these novel alleles were observed in one or two individuals (**Figure 2B**). Seven of them have sequences identical to ones reported recently in a survey of more than one million registered bone marrow donors (47).

On the basis of allele composition, we defined 170 centromeric and 94 telomeric *KIR* haplotypes (**Supplementary Materials B,C**). Thus, by distinct haplotype number, the centromeric region is twice as diverse as the telomeric region. Emphasizing this difference, the 55 most frequent centromeric haplotypes account for 75% of the total haplotypes, whereas only 13 telomeric haplotypes are sufficient to account for 75% of the telomeric haplotypes (**Figures 3A,B**). In conclusion, the centromeric *KIR* region of Iranian *KIR* haplotypes is far more diverse than the telomeric *KIR* region.

Although *KIR-A* haplotypes are more numerous and frequent than *KIR-B* haplotypes (**Figure 1**), the most frequent centromeric region and the second most frequent telomeric region haplotypes are *KIR-B* (**Figures 3C,D**). Together, these centromeric and telomeric segments encode only one inhibitory receptor specific for HLA class I (KIR2DL2*001). Also encoded are KIR2DL1*004, an attenuated receptor (23) and KIR3DL2*007, which has a mutation in the first ITIM of the cytoplasmic tail (48) and thus may not transmit an inhibitory signal. By contrast, the *KIR-B* haplotype encodes three functional activating receptors (KIR2DS1, 2DS2, and 3DS1) and a full-length KIR2DL4 (*005). This combination of common centromeric and telomeric *KIR-B*



A

| | Baloch (N=33) | | Kurds (N=48) | | Lurs | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Lorestan (N=64) | | Lordegan (N=15) | | Yasuj (N=81) | | |
| *KIR gene* | k | H | k | H | k | H | k | H | k | H | Total |
| *2DL1* | 7 | 0.94 | 8 | 0.81 | 8 | 0.85 | 7 | 0.86 | 9 | 0.68 | 12 |
| *2DL2/3* | 5 | 0.76 | 6 | 0.67 | 5 | 0.75 | 5 | 0.73 | 8 | 0.76 | 9 |
| *2DL4* | 8 | 0.79 | 8 | 0.81 | 8 | 0.82 | 8 | 0.93 | 10 | 0.82 | 12 |
| *2DL5* | 7 | 0.63 | 6 | 0.48 | 6 | 0.51 | 3 | 0.53 | 6 | 0.73 | 10 |
| *2DS1* | 1 | 0.42 | 1 | 0.33 | 1 | 0.31 | 1 | 0.4 | 2 | 0.43 | 2 |
| *2DS2* | 1 | 0.36 | 1 | 0.44 | 2 | 0.46 | 1 | 0.73 | 1 | 0.48 | 2 |
| *2DS3/5* | 3 | 0.64 | 3 | 0.60 | 4 | 0.57 | 3 | 0.53 | 5 | 0.70 | 7 |
| *2DS4* | 5 | 0.79 | 5 | 0.62 | 5 | 0.80 | 5 | 0.8 | 6 | 0.76 | 6 |
| *3DL1/S1* | 10 | 0.79 | 15 | 0.83 | 14 | 0.86 | 9 | 0.93 | 15 | 0.85 | 19 |
| *3DL2* | 10 | 0.79 | 16 | 1.00 | 17 | 0.87 | 9 | 0.93 | 17 | 0.81 | 22 |
| *3DL3* | 20 | 0.94 | 26 | 0.96 | 26 | 0.97 | 15 | 0.93 | 39 | 0.99 | 44 |

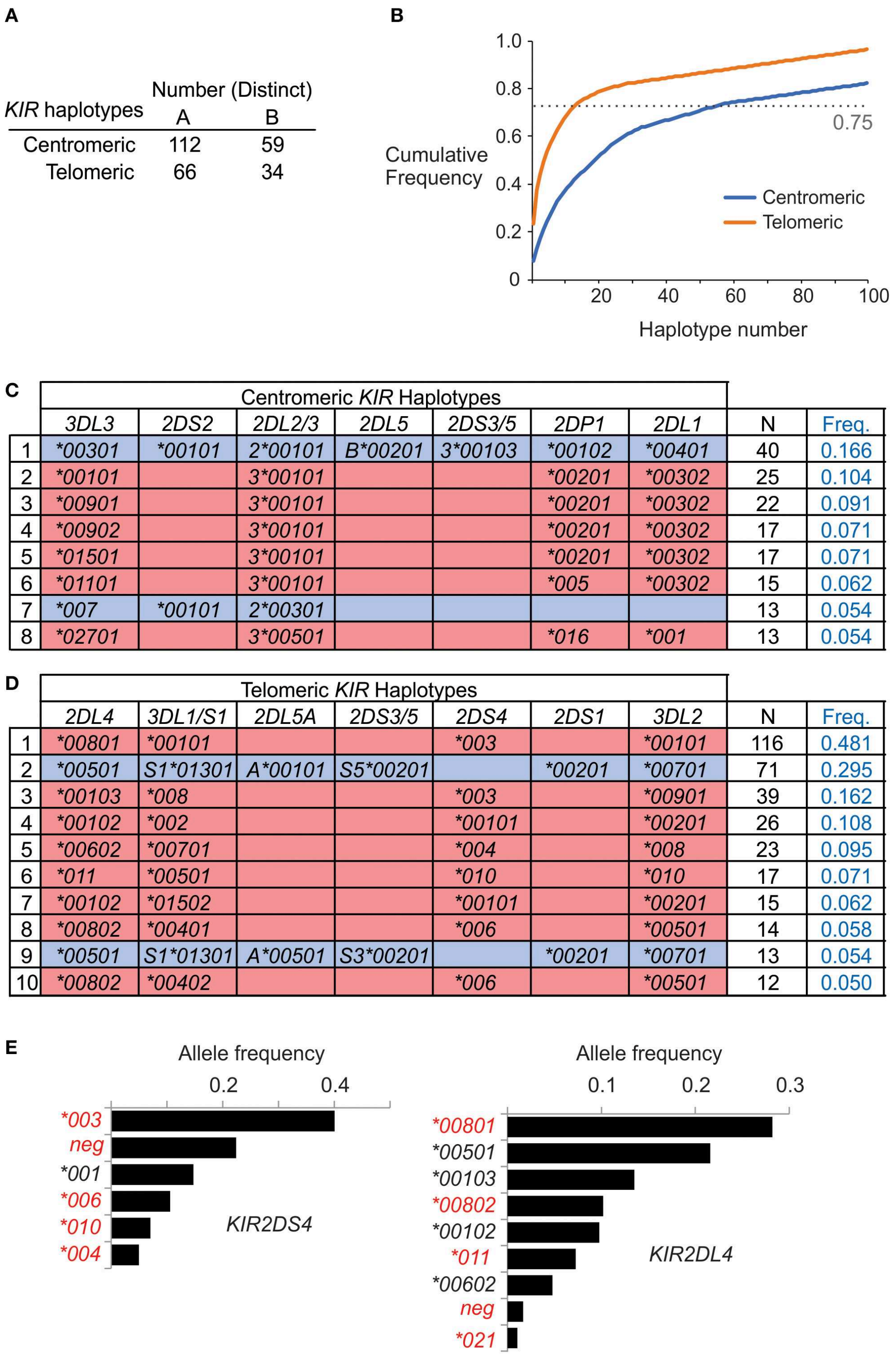
B

| *KIR* gene | Closest allele | Nucleotide change | Coding change | Domain | N |
|---|---|---|---|---|---|
| *2DL1* | *00201 | G 331 T † | V 111 L | D0 | 1 |
| | *00302 | G 421 A † | A 141 T | D1 | 2 |
| *2DL4* | *00801 | G 568 A † | G 190 R | D2 | 2 |
| | *00801 | C 883 A | P 295 T | Cyt | 1 |
| *3DL3* | *01402 | G 102 A † | W 34 X | D0 | 1 |
| *2DS4* | *003 | A 110 C † | H 37 P | D1 | 2 |
| *2DS5* | *00201 | C 252 T † | syn | D0 | 1 |
| *3DS1* | *01301 | G 775 C | G 259 R | D2 | 1 |
| | *01301 | T 1160 C | F 387 L | Cyt | 2 |
| *2DP1* | *00301 | C 702 A † | - | - | |
| | | C 703 G † | - | - | 2 |
| | *00201 | C 1107 T | - | - | 1 |

**FIGURE 2 |** High *KIR* diversity and allele discovery in Iranians. **(A)** Shows the numbers of *KIR* alleles (k) and the heterozygosity (H) in each of the five Iranian populations. The total number observed is shown at the right. Gene absence is not included as an allele. **(B)** Shows the novel *KIR* alleles identified in this study. Columns from left to right are: the *KIR* gene, the closest known allele, the nucleotide change compared to the closest allele, the amino acid substitution caused by the nucleotide change, domain affected by the amino acid substitution (LP, leader peptide; D0–D1, Ig-like domains; TM, transmembrane domain) and the number observed. $^{\dagger}$-indicates identical allele observed by Wagner et al. (47).

haplotypes can therefore provide the maximum number of activating KIR. The most frequent *KIR-A* haplotype encodes four inhibitory receptors specific for polymorphic HLA class I and carries *KIR2DS4*003*, which is not-expressed because of a 22 bp deletion in exon 5 (11). Indeed, the frequency of the *KIR2DS4 22 bp-del* variant in Iranians is 0.62, more than four times higher than the frequency of the full-length variant (**Figure 3E**). Consequently, almost all *KIR-A* haplotypes in Iranians encode four functional inhibitory receptors and no activating receptor specific for polymorphic HLA class I.

A characteristic of the *KIR* locus is the occurrence of large-scale duplication or deletion events that encompass complete genes, which synergizes with allele variation to enhance KIR functional diversity (45, 49, 50). In the Iranian cohort, we identified seven *KIR* haplotypes having large deletions and eight with duplications (**Figure 4**). One of these haplotypes (number 1 in **Figure 4**) was not observed previously and is similar to the most frequent *KIR-B* haplotype in Iran, with the difference being that it lacks *KIR2DS1*. This haplotype could have been formed by homologous or looping out recombination (50). The remaining



A

| *KIR* haplotypes | Number (Distinct) A | B |
|---|---|---|
| Centromeric | 112 | 59 |
| Telomeric | 66 | 34 |

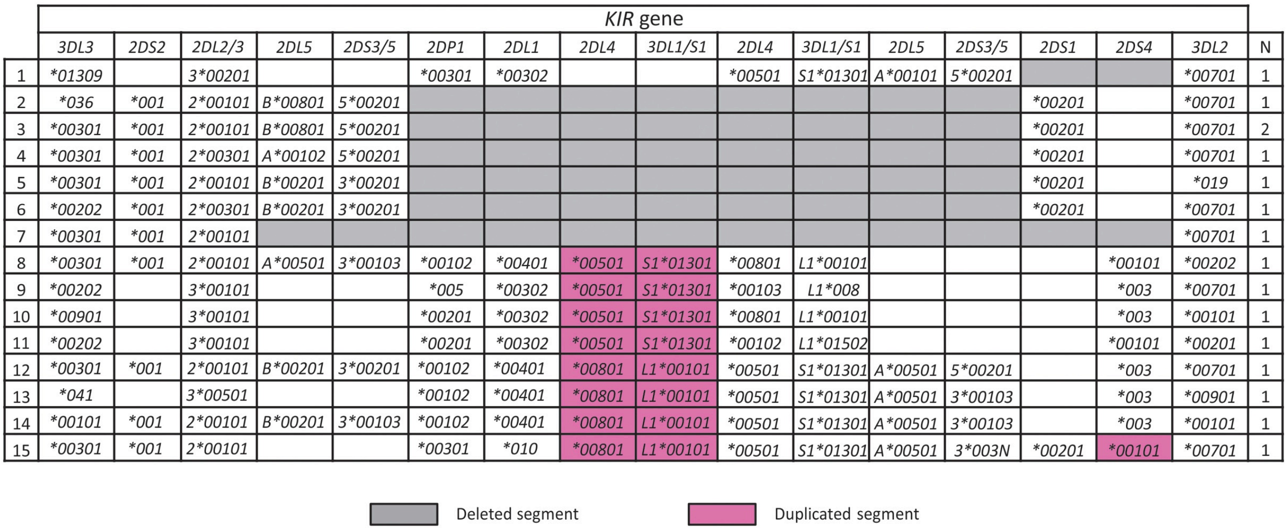
C

| | Centromeric *KIR* Haplotypes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *3DL3* | *2DS2* | *2DL2/3* | *2DL5* | *2DS3/5* | *2DP1* | *2DL1* | N | Freq. |
| 1 🟦 | **00301* | **00101* | *2*00101* | *B*00201* | *3*00103* | **00102* | **00401* | 40 | 0.166 |
| 2 🟥 | **00101* | | *3*00101* | | | **00201* | **00302* | 25 | 0.104 |
| 3 🟥 | **00901* | | *3*00101* | | | **00201* | **00302* | 22 | 0.091 |
| 4 🟥 | **00902* | | *3*00101* | | | **00201* | **00302* | 17 | 0.071 |
| 5 🟥 | **01501* | | *3*00101* | | | **00201* | **00302* | 17 | 0.071 |
| 6 🟥 | **01101* | | *3*00101* | | | **005* | **00302* | 15 | 0.062 |
| 7 🟦 | **007* | **00101* | *2*00301* | | | | | 13 | 0.054 |
| 8 🟥 | **02701* | | *3*00501* | | | **016* | **001* | 13 | 0.054 |

D

| | Telomeric *KIR* Haplotypes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *2DL4* | *3DL1/S1* | *2DL5A* | *2DS3/5* | *2DS4* | *2DS1* | *3DL2* | N | Freq. |
| 1 🟥 | **00801* | **00101* | | | **003* | | **00101* | 116 | 0.481 |
| 2 🟦 | **00501* | *S1*01301* | *A*00101* | *S5*00201* | | **00201* | **00701* | 71 | 0.295 |
| 3 🟥 | **00103* | **008* | | | **003* | | **00901* | 39 | 0.162 |
| 4 🟥 | **00102* | **002* | | | **00101* | | **00201* | 26 | 0.108 |
| 5 🟥 | **00602* | **00701* | | | **004* | | **008* | 23 | 0.095 |
| 6 🟥 | **011* | **00501* | | | **010* | | **010* | 17 | 0.071 |
| 7 🟥 | **00102* | **01502* | | | **00101* | | **00201* | 15 | 0.062 |
| 8 🟥 | **00802* | **00401* | | | **006* | | **00501* | 14 | 0.058 |
| 9 🟦 | **00501* | *S1*01301* | *A*00501* | *S3*00201* | | **00201* | **00701* | 13 | 0.054 |
| 10 🟥 | **00802* | **00402* | | | **006* | | **00501* | 12 | 0.050 |

**FIGURE 3 |** Centromeric *KIR* are more diverse than telomeric *KIR* in Iranians. **(A)** Shows the number of distinct haplotypes observed in the centromeric and telomeric *KIR* regions. **(B)** Shows the cumulative frequency of haplotypes observed in the centromeric and telomeric *KIR* regions. **(C,D)** Shows the allele composition of all **(C)** centromeric, and **(D)** telomeric *KIR* haplotypes identified in more than ten individuals in the combined Iranian population. *KIR A* haplotypes are shaded pink and *KIR B* haplotypes are blue. Empty boxes indicate gene absence. At the right is shown number observed and the frequency in the combined Iranian population (*N* = 241). All the observed haplotypes are given in **Supplementary Material (E)**. Shown are the frequencies of *KIR2DL4* and *KIR2DS4* alleles in the combined Iranian population. Red text indicates alleles not expressed at cell surface. The allele frequencies for all *KIR* genes in each of the five populations are given in **Supplementary Material**.



| | *KIR* gene | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3DL3 | 2DS2 | 2DL2/3 | 2DL5 | 2DS3/5 | 2DP1 | 2DL1 | 2DL4 | 3DL1/S1 | 2DL4 | 3DL1/S1 | 2DL5 | 2DS3/5 | 2DS1 | 2DS4 | 3DL2 | N |
| 1 | *01309 | | 3*00201 | | | *00301 | *00302 | | | *00501 | S1*01301 | A*00101 | 5*00201 | ▨ | ▨ | *00701 | 1 |
| 2 | *036 | *001 | 2*00101 | B*00801 | 5*00201 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | *00201 | | *00701 | 1 |
| 3 | *00301 | *001 | 2*00101 | B*00801 | 5*00201 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | *00201 | | *00701 | 2 |
| 4 | *00301 | *001 | 2*00301 | A*00102 | 5*00201 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | *00201 | | *00701 | 1 |
| 5 | *00301 | *001 | 2*00101 | B*00201 | 3*00201 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | *00201 | | *019 | 1 |
| 6 | *00202 | *001 | 2*00301 | B*00201 | 3*00201 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | *00201 | | *00701 | 1 |
| 7 | *00301 | *001 | 2*00101 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | *00701 | 1 |
| 8 | *00301 | *001 | 2*00101 | A*00501 | 3*00103 | *00102 | *00401 | 🟪 *00501 | 🟪 S1*01301 | *00801 | L1*00101 | | | | *00101 | *00202 | 1 |
| 9 | *00202 | | 3*00101 | | | *005 | *00302 | 🟪 *00501 | 🟪 S1*01301 | *00103 | L1*008 | | | | *003 | *00701 | 1 |
| 10 | *00901 | | 3*00101 | | | *00201 | *00302 | 🟪 *00501 | 🟪 S1*01301 | *00801 | L1*00101 | | | | *003 | *00101 | 1 |
| 11 | *00202 | | 3*00101 | | | *00201 | *00302 | 🟪 *00501 | 🟪 S1*01301 | *00102 | L1*01502 | | | | *00101 | *00201 | 1 |
| 12 | *00301 | *001 | 2*00101 | B*00201 | 3*00201 | *00102 | *00401 | 🟪 *00801 | 🟪 L1*00101 | *00501 | S1*01301 | A*00501 | 5*00201 | | *003 | *00701 | 1 |
| 13 | *041 | | 3*00501 | | | *00102 | *00401 | 🟪 *00801 | 🟪 L1*00101 | *00501 | S1*01301 | A*00501 | 3*00103 | | *003 | *00901 | 1 |
| 14 | *00101 | *001 | 2*00101 | B*00201 | 3*00103 | *00102 | *00401 | 🟪 *00801 | 🟪 L1*00101 | *00501 | S1*01301 | A*00501 | 3*00103 | | *003 | *00101 | 1 |
| 15 | *00301 | *001 | 2*00101 | | | *00301 | *010 | 🟪 *00801 | 🟪 L1*00101 | *00501 | S1*01301 | A*00501 | 3*003N | *00201 | 🟪 *00101 | *00701 | 1 |

▨ Deleted segment

🟪 Duplicated segment

**FIGURE 4 |** Rare structural variants of KIR haplotypes in Iranians. Shown are Iranian *KIR* haplotypes affected by a copy number variation. Gray and purple boxes highlight deleted or duplicated segments, respectively. "*N*" indicates the number observed.

deletion haplotypes are similar to those observed worldwide including Africans, and the duplication haplotypes similar to those observed outside of Africa (45).

The interaction of HLA class I with inhibitory KIR contributes to NK cell education (51, 52). We analyzed the distribution of alleles for the four inhibitory KIR that are specific for HLA class I in Iranians and compared them with six other populations that represent the breadth of human genetic diversity (**Figure 5**). This analysis showed that in Iranians, Europeans, West Africans, and Japanese the most frequent *KIR2DL1* and *KIR2DL2/3* alleles are *2DL1*003* and *2DL3*001*, respectively. The two populations that differ are relatively small indigenous populations (**Figures 5A,B**). By contrast, the *KIR3DL1/S1* and *KIR3DL2* alleles most frequent in Iranians are usually not the same as those most frequent in other populations (**Figures 5C,D**). This shows there is a greater worldwide divergence of KIR specific for HLA-A and -B than of KIR specific for HLA-C. Of particular note, Iranian populations have the highest frequency of *3DL1*001* (0.29) compared to the other six populations as well as to all other populations analyzed to date (53). Allele frequencies of the four inhibitory *KIR* specific for HLA class I are similar across the Iranian populations analyzed (**Figures 5E–H**). The one exception is the Baloch who have a low frequency of *KIR3DL1*004*, as well as *KIR3DL2*003* and **005,* which are in strong linkage disequilibrium with *3DL1*004* (28, 44) (**Supplementary Material C**). The Baloch have one tenth the frequency of *3DL1*004* (0.015) than Kurd [0.114, $\chi 2$ $p = 0.006$; pc = 0.06(ns)], Lorestan [0.109, $\chi 2$ $p = 0.006$; pc = 0.06(ns)] and Yasuj [0.129, $\chi 2$ $p = 0.003$; pc = 0.03]. KIR3DL1*004 is retained in the cytoplasm and unable to bind HLA-Bw4+ HLA-A or -B on target cells (54).

We examined the compound genotypes of *KIR* and *HLA class I*, to determine the potential number of interactions between HLA class I ligands and inhibitory KIR. We observed a consistent mean of four viable interactions per individual of HLA-C with inhibitory KIR across the five Iranian populations (**Figure 6A**). This number is similar to the mean of 3.6 observed in Europeans (28). Although we observed no significant difference in the potential interactions of KIR3DL1 with HLA-A, the Baloch and Lordegan have more interactions of KIR3DL1 with HLA-B than the other three Iranian populations (**Figures 6B,C**). Despite the small numbers of individuals in these groups, the differences are statistically significant (Mann-Whitney *U*-test; $p < 0.013$). Contributing to these differences is the high frequency of HLA-B*51 in the Baloch (0.29; **Supplementary Material D**). This frequency is similar to the 0.28 observed in the Lordegan and considered the highest worldwide (34).

## DISCUSSION

This study applied high-throughput, next-generation sequencing to define *KIR* polymorphism at high resolution in five Iranian populations. These comprise the Kurd and Lur populations from the Zagros Mountains in the west of Iran and the Baloch from the south-east. The Lur comprised three subpopulations; the Lorestan, Lordegan, and Yasuj. *HLA class I* allele distributions for four of these populations were reported previously (34) whereas those for the Baloch are described here. When we compared the *KIR* allele frequencies of Iranians to those representing African, Asian, European, Oceanian, and South American populations, we found that the Iranian groups we studied are particularly similar to Europeans. This finding is consistent with ancient DNA analysis, which revealed that Iranians and Europeans both originate from an Indo-Europeans steppe ancestor population (55).

Despite their overall similarity, the Iranians we studied differ from Europeans in their frequencies of *KIR3DL1*001*, a high-expressing inhibitory receptor specific for the Bw4 epitope of

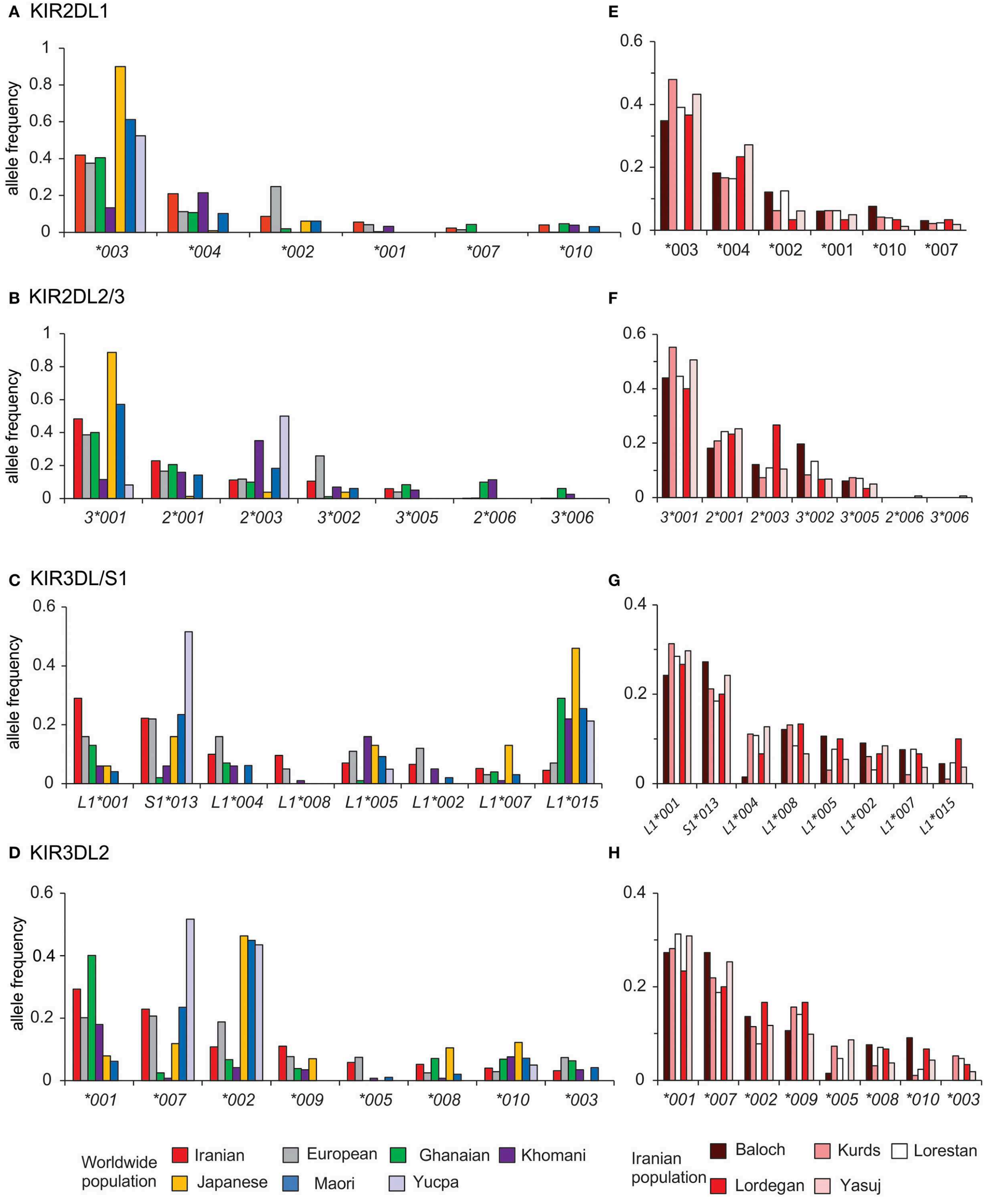


**FIGURE 5 |** Telomeric *KIR* are more divergent than centromeric *KIR* across populations. Shows the alleles of **(A)** *KIR2DL1*, **(B)** *KIR2DL2/3*, **(C)** *KIR3DL1/S1*, and **(D)** *KIR3DL2* observed in Iran, and their frequencies in the combined Iranian population and six other populations representing major world groups (Left). **(E-H)** show the allele frequencies in the individual Iranian populations. The populations are Iranians (*N* = 241, comprised from Baloch *N* = 33, Kurds *N* = 48, Lorestan *N* = 64, Lordegan *N* = 15, and Yasuj *N* = 81), Ghanaians (*N* = 131), Khomani (*N* = 79), Maori (*N* = 49), Japanese (*N* = 115), Europeans (*N* = 378), and Yucpa (*N* = 61). The allele frequencies for all *KIR* genes in each of the five Iranian populations are given in **Supplementary Material**.

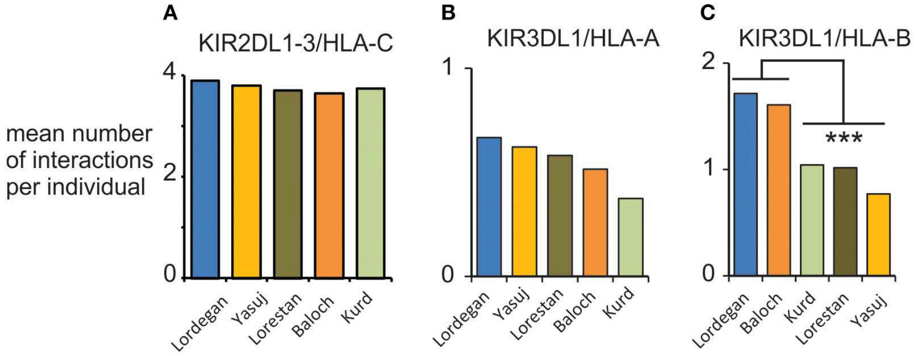


**FIGURE 6 |** Interactions between KIR and HLA class I in five Iranian populations. **(A)** Shows the mean total number of viable interactions per individual of inhibitory KIR and HLA-C. **(B)** Shows the mean total number of viable interactions per individual of KIR3DL1 and HLA-A. **(C)** Shows the mean total number of viable interactions per individual of KIR3DL1 and HLA-B. ***$p < 0.013$, obtained from a two-tailed Mann-Whitney *U*-test.

subsets of HLA-A and -B allotypes. Iranians have *KIR3DL1*001* frequencies that are twice those in Europeans and are the highest worldwide. Iran has the highest frequency of *HLA-B*51* in the world (34), and this is also the case for the Baloch. HLA-B*51 has the Bw4 epitope and thus educates KIR3DL1$^+$ NK cells to detect any loss in Bw4$^+$ HLA-A or -B expression (35, 56, 57). HLA-B*51 is associated with Behçet disease, a chronic, multi-system autoimmune condition that has substantially higher incidence in Iran (80/100,000) than Europe (<1/100,000) (58, 59). We recently showed that high-expressing allotypes of KIR3DL1, including 3DL1*001, can protect from Behçet disease (60). It is unlikely that KIR3DL1*001 and HLA-B*51 rose to high frequency in Iran to protect specifically from an autoimmune disease, but this combination of HLA and KIR could also protect against specific infectious diseases (7). Examples include tuberculosis and hepatitis, which are both prevalent in Balochistan (61–63). Amongst Iranians, the Baloch and Lordegan have the highest number of viable interactions between KIR3DL1 and Bw4$^+$HLA. Contributing to this high occurrence in the Baloch, are high frequencies of KIR3DL1*001 and HLA-B*51 and a low frequency of KIR3DL1*004. That both the *KIR3DL2* alleles linked to *KIR3DL1*004* are also reduced in frequency suggests that KIR3DL1*004 has been specifically targeted by negative selection in the Baloch, rather than another *KIR* allele in linkage disequilibrium with *KIR3DL1*004*.

*KIR-A* and *-B* haplotypes are present in all human populations, where they are maintained by balancing selection, likely because *KIR-A* haplotypes favor infection control, particularly viral infections, and *KIR-B* haplotypes favor successful fetal implantation (1, 16). Accordingly, *KIR-A* haplotypes express all possible inhibitory receptors specific for HLA class I, whereas *KIR-B* haplotypes express fewer inhibitory receptors but more activating receptors. Thus, whereas the inhibitory KIR help prime NK cells to be able to be responsive to HLA class I loss during infection, the activating KIR can both promote fetal trophoblast invasion and recognize specific pathogen-derived peptides to control certain infections (64, 65). In Iranian populations the differences between the *KIR-A* and *KIR-B* haplotypes is extreme. The common *KIR-A* haplotypes express no activating KIR, due to the high frequency of the truncated KIR2DS4 variant, and the common *KIR-B* haplotypes provide the maximum number of activating receptors. In this regard, the *KIR-B* haplotypes differ considerably from those of Africans and were likely obtained since the out of Africa migration, through adaptive introgression with ancient humans (66).

In summary, we describe the *KIR* locus at allelic resolution in Iranian populations and place it in the context of the HLA ligands recognized by KIR. Iran is a culturally diverse country and the ethnic groups we have studied comprise ~15% of the population (67). Because substantial *KIR* gene content diversity is observed across Iran (68–71), it will be of interest in future studies to compare our results with other Iranian populations including Persians and Azeris. The allele and haplotype distributions described here will provide a baseline for future studies of disease association and transplantation matching in this important region of the world.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the article/**Supplementary Material**. Other raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Research Ethics Committee of Shiraz University of Medical Sciences. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

PP, EA, AG, and PN conceived and designed experiments. EA, NN-G, and ST performed lab experiments. CA, PN, JAT, LG, NN-G, and ST analyzed data. AG, PP, WM, JH, JT, and LM provided materials. CA, PN, and PP wrote the paper. All authors approved the final submitted version.

## FUNDING

This work was supported by National Institutes of Health Grants R01 AI017892 and U01 AI090905. The project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 695551).

## ACKNOWLEDGMENTS

We thank all of the participants of this study for their generous donation of DNA to facilitate genetic research. CA is recipient of a post-doctoral fellowship from the Fondazione Umberto Veronesi.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00556/full#supplementary-material

**Supplementary Material (Spreadsheet) | (A)** Shown are the allele frequencies for the 13 *KIR* genes in five Iranian populations, and the combined frequency is shown at the right. **(B)** Shown are the *centromeric KIR* haplotypes identified in five Iranian populations. The combined frequency is shown at the right. *KIR A* haplotypes are shaded in red, *KIR B* haplotypes are shaded in blue. **(C)** Shown are the *telomeric KIR* haplotypes identified in five Iranian populations. The combined frequency is shown at the right. *KIR A* haplotypes are shaded in red, *KIR B* haplotypes are shaded in blue. Yellow shading indicates novel *KIR* alleles identified in this study, which are described in **Figure 2**. **(D)** Shows the *HLA class I* allele frequencies observed in the Baloch population.

## REFERENCES

1. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol.* (2013) 13:133–44. doi: 10.1038/nri3370
2. Vance RE, Eichberg MJ, Portnoy DA, Raule, D. Listening to each other: infectious disease and cancer immunology. *Sci Immunol.* (2017) 2:eaai9339. doi: 10.1126/sciimmunol.aai9339
3. Long EO, Kim HS, Liu D, Peterson ME, Rajagopalan S. Controlling natural killer cell responses: integration of signals for activation and inhibition. *Annu Rev Immunol.* (2013) 31:227–58. doi: 10.1146/annurev-immunol-020711-075005
4. Lam VC, Lanier LL. NK cells in host responses to viral infections. *Curr Opin Immunol.* (2017) 44:43–51. doi: 10.1016/j.coi.2016.11.003
5. Stewart CA, Vivier E, Colonna M. Strategies of natural killer cell recognition and signaling. *Curr Top Microbiol Immunol.* (2006) 298:1–21. doi: 10.1007/3-540-27743-9_1
6. Moretta L, Pietra G, Montaldo E, Vacca P, Pende D, Falco M, et al. Human NK cells: from surface receptors to the therapy of leukemias and solid tumors. *Front Immunol.* (2014) 5:87. doi: 10.3389/fimmu.2014.00087
7. Bashirova AA, Martin MP, McVicar, DW, Carrington M. The killer immunoglobulin-like receptor gene cluster: tuning the genome for defense. *Annu Rev Genomics Hum Genet.* (2006) 7:277–300. doi: 10.1146/annurev.genom.7.080505.115726
8. Locatelli F, Pende D, Falco M, Della Chiesa M, Moretta A, Moretta L. NK cells mediate a crucial graft-versus-leukemia effect in haploidentical-HSCT to cure high-risk acute leukemia. *Trends Immunol.* (2018) 39:577–90. doi: 10.1016/j.it.2018.04.009
9. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci USA.* (2000) 97:4778–83. doi: 10.1073/pnas.080588597
10. Uhrberg M, Valiante NM, Shum BP, Shilling H, Lienert-Weidenbach K, Corliss B, et al. Human diversity in killer cell inhibitory receptor genes. *Immunity.* (1997) 7:753–63. doi: 10.1016/S1074-7613(00)80394-5
11. Hsu KC, Liu XR, Selvakumar A, Mickelson E, O'Reilly R, Dupont B. Killer Ig-like receptor haplotype analysis by gene content: evidence for genomic diversity with a minimum of six basic framework haplotypes, each with multiple subsets. *J Immunol.* (2002) 169:5118–29. doi: 10.4049/jimmunol.169.9.5118
12. Marsh SG, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Tissue Antigens.* (2003) 62:79–86. doi: 10.1034/j.1399-0039.2003.00072.x
13. Kikuchi-Maki A, Catina TL, Campbell KS. Cutting edge: KIR2DL4 transduces signals into human NK cells through association with the Fc receptor gamma protein. *J Immunol.* (2005) 174:3859–63. doi: 10.4049/jimmunol.174.7.3859
14. Miah SM, Hughes TL, Campbell KS. KIR2DL4 differentially signals downstream functions in human NK cells through distinct structural modules. *J Immunol.* (2008) 180:2922–32. doi: 10.4049/jimmunol.180.5.2922
15. Rajagopalan S, Long EO. KIR2DL4 (CD158d): an activation receptor for HLA-G. *Front Immunol.* (2012) 3:258. doi: 10.3389/fimmu.2012.00258
16. Guethlein LA, Norman PJ, Hilton HH, Parham P. Co-evolution of MHC class I and variable NK cell receptors in placental mammals. *Immunol Rev.* (2015) 267:259–82. doi: 10.1111/imr.12326
17. Pyo CW, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different patterns of evolution in the centromeric and telomeric regions of group A and B haplotypes of the human killer cell Ig-like receptor locus. *PLoS ONE.* (2010) 5:e15115. doi: 10.1371/journal.pone.0015115
18. Cooley S, Weisdorf DJ, Guethlein LA, Klein J, Wang T, Le CT, et al. Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia. *Blood.* (2010) 116:2411–9. doi: 10.1182/blood-2010-05-283051
19. Oevermann L, Michaelis SU, Mezger M, Lang P, Toporski J, Bertaina A, et al. KIR B haplotype donors confer a reduced risk for relapse after haploidentical transplantation in children with all. *Blood.* (2014) 124:2744–7. doi: 10.1182/blood-2014-03-565069
20. Winter CC, Gumperz JE, Parham P, Long EO, Wagtmann N. Direct binding and functional transfer of NK cell inhibitory receptors reveal novel patterns of HLA-C allotype recognition. *J Immunol.* (1998) 161:571–7.
21. Pando MJ, Gardiner CM, Gleimer M, McQueen KL, Parham P. The protein made from a common allele of KIR3DL1 (3DL1*004) is poorly expressed at cell surfaces due to substitution at positions 86 in Ig domain 0 182 in Ig domain 1. *J Immunol.* (2003) 171:6640–9. doi: 10.4049/jimmunol.171.12.6640
22. Moesta AK, Norman PJ, Yawata M, Yawata N, Gleimer M, Parham P. Synergistic polymorphism at two positions distal to the ligand-binding site makes KIR2DL2 a stronger receptor for HLA-C than KIR2DL3. *J Immunol.* (2008) 180:3969–79. doi: 10.4049/jimmunol.180.6.3969
23. Bari R, Bell T, Leung WH, Vong QP, Chan WK, Das Gupta N. Significant functional heterogeneity among KIR2DL1 alleles and a pivotal role of arginine 245. *Blood.* (2009) 114:5182–90. doi: 10.1182/blood-2009-07-231977
24. VandenBussche CJ, Mulrooney TJ, Frazier WR, Dakshanamurthy S, Hurley CK. Dramatically reduced surface expression of NK cell receptor KIR2DS3 is attributed to multiple residues throughout the molecule. *Genes Immun.* (2009) 10:162–73. doi: 10.1038/gene.2008.91
25. Hilton HG, Guethlein LA, Goyos A, Nemat-Gorgani N, Bushnell D, Norman PJ, et al. Polymorphic HLA-C receptors balance the functional characteristics of KIR haplotypes. *J Immunol.* (2015) 195:3160–70. doi: 10.4049/jimmunol.1501358
26. Gendzekhadze K, Norman PJ, Abi-Rached L, Graef T, Moesta A, Layrisse Z, et al. Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. *Proc Natl Acad Sci USA.* (2009) 106:18692–7. doi: 10.1073/pnas.0906051106
27. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Guethlein LA, Hilton HG, Pando MJ, et al. Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. *PLoS Genet.* (2013) 9:e1003938. doi: 10.1371/journal.pgen.1003938

28. Nemat-Gorgani N, Hilton HG, Henn BM, Lin M, Gignoux CR, Myrick JW, et al. Different selected mechanisms attenuated the inhibitory interaction of KIR2DL1 with C2(+) HLA-C in two indigenous human populations in Southern Africa. *J Immunol.* (2018) 200:2640–55. doi: 10.4049/jimmunol.1701780
29. Nemat-Gorgani N, Guethlein LA, Henn BM, Norberg S, Chiaroni J, Sikora M, et al. Diversity of KIR, HLA class I, their interactions in seven populations of sub-Saharan Africans. *J Immunol.* (2019) 202:2636–47. doi: 10.4049/jimmunol.1801586
30. Middleton D, Meenagh A, Gourraud PA. KIR haplotype content at the allele level in 77 Northern Irish families. *Immunogenetics.* (2007) 59:145–58. doi: 10.1007/s00251-006-0181-7
31. Vierra-Green C, Roe D, Hou L, Hurley CK, Rajalingam R, Reed E, et al. Allele-level haplotype frequencies and pairwise linkage disequilibrium for 14 KIR loci in 506 European-American individuals. *PLoS ONE*. (2012) 7:e47491. doi: 10.1371/journal.pone.0047491
32. Yawata M, Yawata N, Draghi M, Little AM, Partheniou F, Parham P. Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. *J Exp Med.* (2006) 203:633–45. doi: 10.1084/jem.20051884
33. Nemat-Gorgani N, Edinur HA, Hollenbach JA, Traherne J, Dunn PP, Chambers GK, et al. KIR diversity in Maori and Polynesians: populations in which HLA-B is not a significant KIR ligand. *Immunogenetics.* (2014) 66:597–611. doi: 10.1007/s00251-014-0794-1
34. Ashouri E, Norman PJ, Guethlein LA, Han A, Nemat-Gorgani N, Norberg SJ, et al. HLA class I variation in Iranian Lur and Kurd populations: high haplotype and allotype diversity with an abundance of KIR ligands. *HLA*. (2016) 88:87–99. doi: 10.1111/tan.12852
35. Cella M, Longo A, Ferrara GB, Strominger JL, Colonna M. NK3-specific natural killer cells are selectively inhibited by Bw4-positive HLA alleles with isoleucine 80. *J Exp Med.* (1994) 180:1235–42. doi: 10.1084/jem.180.4.1235
36. Saunders PM, Pymm P, Pietra G, Hughes VA, Hitchen C, O'Connor GM, et al. Killer cell immunoglobulin-like receptor 3DL1 polymorphism defines distinct hierarchies of HLA class I recognition. *J Exp Med.* (2016) 213:791–807. doi: 10.1084/jem.20152023
37. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. *Am J Hum Genet.* (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023
38. Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. (2012) 9:357–9. doi: 10.1038/nmeth.1923
39. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller W, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* (2004) 14:1147–59. doi: 10.1101/gr.1917404
40. Bonfield JK, Whitwham A. Gap5–editing the billion fragment sequence assembly. *Bioinformatics*. (2010) 26:1699–703. doi: 10.1093/bioinformatics/btq268
41. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics*. (2013) 14:178–92. doi: 10.1093/bib/bbs017
42. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* (2003) 73:1162–9. doi: 10.1086/379378
43. Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis. models and estimation procedures. *Am J Hum Genet.* (1967) 19(3 Pt 1):233–57. doi: 10.2307/2406616
44. Misra MK, Augusto DG, Martin GM, Nemat-Gorgani N, Sauter J, Hofmann JA, et al. Report from the killer-cell immunoglobulin-like receptors (KIR) component of the 17th international hla and immunogenetics workshop. *Hum Immunol.* (2018) 79:825–33. doi: 10.1016/j.humimm.2018.10.003
45. Norman PJ, Abi-Rached L, Gendzekhadze K, Hammond JA, Moesta A, Sharma D, et al. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. *Genome Res.* (2009) 19:757–69. doi: 10.1101/gr.085738.108
46. Hollenbach JA, Nocedal I, Ladner MB, Single RM, Trachtenberg EA. Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations. *Immunogenetics*. (2012) 64:719–37. doi: 10.1007/s00251-012-0629-x
47. Wagner I, Schefzyk D, Pruschke J, Schofl G, Schone B, Gruber N, et al. Allele-level KIR genotyping of more than a million samples: workflow, algorithm, and observations. *Front Immunol.* (2018) 9:2843. doi: 10.3389/fimmu.2018.02843
48. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG, et al. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. (2015) 43:D423–31. doi: 10.1093/nar/gku1161
49. Martin MP, Bashirova A, Traherne J, Trowsdale J, Carrington M. Cutting edge: expansion of the KIR locus by unequal crossing over. *J Immunol.* (2003) 171:2192–5. doi: 10.4049/jimmunol.171.5.2192
50. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet.* (2010) 19:737–51. doi: 10.1093/hmg/ddp538
51. Anfossi N, Andre P, Guia S, Falk CS, Roetynck S, Stewart CA, et al. Human NK cell education by inhibitory receptors for MHC class I. *Immunity.* (2006) 25:331–42. doi: 10.1016/j.immuni.2006.06.013
52. Goodridge JP, Jacobs B, Saetersmoen ML, Clement D, Hammer Q, Clancy T, et al. Remodeling of secretory lysosomes during education tunes functional potential in NK cells. *Nat Commun.* (2019) 10:514. doi: 10.1038/s41467-019-08384-x
53. Norman PJ, Abi-Rached L, Gendzekhadze K, Korbel D, Gleimer M, Rowley D, et al. Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans. *Nat Genet.* (2007) 39:1092–9. doi: 10.1038/ng2111
54. Taner SB, Pando MJ, Roberts A, Schellekens J, Marsh S, Malmberg KJ, et al. Interactions of NK cell receptor KIR3DL1*004 with chaperones and conformation-specific antibody reveal a functional folded state as well as predominant intracellular retention. *J Immunol.* (2011) 186:62–72. doi: 10.4049/jimmunol.0903657
55. Broushaki F, Thomas MG, Link V, Lopez S, van Dorp L, Kirsanow K, et al. Early neolithic genomes from the eastern Fertile Crescent. *Science*. (2016) 353:499–503. doi: 10.1126/science.aaf7943
56. Gumperz JE, Litwin V, Phillips JH, Lanier LL, Parham P. The Bw4 public epitope of HLA-B molecules confers reactivity with natural killer cell clones that express NKB1, a putative HLA receptor. *J Exp Med.* (1995) 181:1133–44. doi: 10.1084/jem.181.3.1133
57. Carr WH, Pando MJ, Parham P. KIR3DL1 polymorphisms that affect NK cell inhibition by HLA-Bw4 ligand. *J Immunol.* (2005) 175:5222–9. doi: 10.4049/jimmunol.175.8.5222
58. Ohno S, Aoki K, Sugiura S, Nakayama E, Itakura K, Aizawa M. Letter: HL-A5 and Behcet's disease. *Lancet*. (1973) 2:1383–4. doi: 10.1016/S0140-6736(73)93343-6
59. Verity DH, Wallace GR, Vaughan RW, Stanford M. Behcet's disease: from Hippocrates to the third millennium. *Br J Ophthalmol.* (2003) 87:1175–83. doi: 10.1136/bjo.87.9.1175
60. Petrushkin H, Norman PJ, Lougee E, Parham P, Wallace G, Stanford MR, et al. KIR3DL1/S1 allotypes contribute differentially to the development of behcet disease. *J Immunol.* (2019) 203:1629–35. doi: 10.4049/jimmunol.1801178
61. Sheikh NS, Sheikh AS, Sheikh AA, Yahya S, Rafi-U-Shan, Lateef M. Sero-prevalence of hepatitis B virus infection in Balochistan Province of Pakistan. *Saudi J Gastroenterol.* (2011) 17:180–4. doi: 10.4103/1319-3767.80380
62. Rahimi Foroushani A, Farzianpour F, Tavana A, Rasouli J, Hosseini S. The 10-year trend of TB rate in West Azerbaijan Province, Iran from 2001 to 2010. *Iran J Public Health.* (2014) 43:778–86.
63. WHO. *Global Tuberculosis Report 2019*. Geneva: World Health Organization (2019).
64. Mbiribindi B, Mukherjee S, Wellington D, Das J, Khakoo SI. Spatial clustering of receptors and signaling molecules regulates NK cell response to peptide repertoire changes. *Front Immunol.* (2019) 10:605. doi: 10.3389/fimmu.2019.02370
65. Sim MJW, Rajagopalan S, Altmann DM, Boyton RJ, Sun PD, Long EO. Human NK cell receptor KIR2DS4 detects a conserved bacterial epitope presented by HLA-C. *Proc Natl Acad Sci USA*. (2019) 116:201903781. doi: 10.1073/pnas.1903781116
66. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science.* (2011) 334:89–94. doi: 10.1126/science.1209202

67. Mehrjoo Z, Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Ardalani F, et al. Distinct genetic variation and heterogeneity of the Iranian population. *PLoS Genet.* (2019) 15:e1008385. doi: 10.1371/journal.pgen.1008385

68. Ashouri E, Farjadian S, Reed EF, Ghaderi A, Rajalingam R. KIR gene content diversity in four Iranian populations. *Immunogenetics.* (2009) 61:483–92. doi: 10.1007/s00251-009-0378-7

69. Tajik N, Shahsavar F, Mousavi T, Radjabzadeh MF. Distribution of KIR genes in the Iranian population. *Tissue Antigens*. (2009) 74:22–31. doi: 10.1111/j.1399-0039.2009.01263.x

70. Hiby SE, Ashrafian-Bonab M, Farrell L, Single RM, Balloux F, Carrington M, et al. Distribution of killer cell immunoglobulin-like receptors (KIR) and their HLA-C ligands in two Iranian populations. *Immunogenetics.* (2010) 62:65–73. doi: 10.1007/s00251-009-0408-5

71. Solgi G, Ghafari H, Ashouri E, Alimoghdam K, Rajalingam R, Amirzargar A. (2011). Comparison of KIR gene content profiles revealed a difference between northern and southern Persians in the distribution of KIR2DS5 and its linked loci. *Hum Immunol.* 72:1079–83. doi: 10.1016/j.humimm.2011.08.002

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 29 May 2020
doi: 10.3389/fimmu.2020.00941

# Capturing Differential Allele-Level Expression and Genotypes of All Classical HLA Loci and Haplotypes by a New Capture RNA-Seq Method

*Fumiko Yamamoto [1,2†], Shingo Suzuki [2†], Akiko Mizutani [2,3], Atsuko Shigenari [2], Sayaka Ito [2], Yoshie Kametani [2], Shunichi Kato [4], Marcelo Fernandez-Viña [1,5], Makoto Murata [6], Satoko Morishima [7], Yasuo Morishima [8], Masafumi Tanaka [2], Jerzy K. Kulski [2,9], Seiamak Bahram [10] and Takashi Shiina [2*]*

*[1] Department of Pathology, Stanford University School of Medicine, Palo Alto, CA, United States, [2] Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan, [3] Faculty of Health and Medical Science, Teikyo Heisei University, Toshima-ku, Tokyo, Japan, [4] Division of Hematopoietic Cell Transplantation, Department of Innovative Medical Science, Tokai University School of Medicine, Isehara, Japan, [5] Histocompatibility, Immunogenetics, and Disease Profiling Laboratory, Stanford Blood Center, Stanford Health Care, Palo Alto, CA, United States, [6] Department of Hematology and Oncology, Nagoya University Graduate School of Medicine, Nagoya, Japan, [7] Division of Endocrinology, Diabetes and Metabolism, Hematology, Rheumatology, Second Department of Internal Medicine, Graduate School of Medicine, University of the Ryukyus, Nishihara, Japan, [8] Department of Promotion for Blood and Marrow Transplantation, Aichi Medical University School of Medicine, Nagakute, Japan, [9] Faculty of Health and Medical Sciences, The University of Western Australia Medical School, Crawley, WA, Australia, [10] Laboratoire d'ImmunoRhumatologie Moléculaire, Plateforme GENOMAX, INSERM UMR_S 1109, LabEx TRANSPLANTEX, Fédération Hospitalo-Universitaire OMICARE, Laboratoire International Associé INSERM FJ-HLA-Japan, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Faculté de Médecine, Université de Strasbourg, Service d'Immunologie Biologique, Nouvel Hôpital Civil, Strasbourg, France*

The highly polymorphic human major histocompatibility complex (MHC) also known as the human leukocyte antigen (HLA) encodes class I and II genes that are the cornerstone of the adaptive immune system. Their unique diversity (>25,000 alleles) might affect the outcome of any transplant, infection, and susceptibility to autoimmune diseases. The recent rapid development of new next-generation sequencing (NGS) methods provides the opportunity to study the influence/correlation of this high level of HLA diversity on allele expression levels in health and disease. Here, we describe the NGS capture RNA-Seq method that we developed for genotyping all 12 classical HLA loci (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5*) and assessing their allelic imbalance by quantifying their allele RNA levels. This is a target enrichment method where total RNA is converted to a sequencing-ready complementary DNA (cDNA) library and hybridized to a complex pool of RNA-specific HLA biotinylated oligonucleotide capture probes, prior to NGS. This method was applied to 161 peripheral blood mononuclear cells and 48 umbilical cord blood cells of healthy donors. The differential allelic expression of 10 HLA loci (except for *HLA-DRA* and *HLA-DPA1*) showed strong significant differences ($P < 2.1 \times 10^{-15}$). The results were corroborated by independent methods. This newly developed NGS method could be applied to a wide range of biological and medical questions including graft rejections and HLA-related diseases.

**Keywords: human leukocyte antigen, next-generation sequencing, HLA allele, RNA expression level, genotyping, capture RNA-Seq**

OPEN ACCESS

***Edited by:***
*Martin Maiers,*
*National Marrow Donor Program,*
*United States*

***Reviewed by:***
*Pietro Crivello,*
*Essen University Hospital, Germany*
*Marco Andreani,*
*Bambino Gesù Ospedale Pediatrico*
*(IRCCS), Italy*

****Correspondence:***
*Takashi Shiina*
*tshiina@is.icc.u-tokai.ac.jp*

*† These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to*
*Alloimmunity and Transplantation,*
*a section of the journal*
*Frontiers in Immunology*

***Received:*** *03 December 2019*
***Accepted:*** *22 April 2020*
***Published:*** *29 May 2020*

***Citation:***
*Yamamoto F, Suzuki S, Mizutani A, Shigenari A, Ito S, Kametani Y, Kato S, Fernandez-Viña M, Murata M, Morishima S, Morishima Y, Tanaka M, Kulski JK, Bahram S and Shiina T (2020) Capturing Differential Allele-Level Expression and Genotypes of All Classical HLA Loci and Haplotypes by a New Capture RNA-Seq Method.*
*Front. Immunol. 11:941.*
*doi: 10.3389/fimmu.2020.00941*

# INTRODUCTION

The highly polymorphic human major histocompatibility complex (MHC), also known as the human leukocyte antigen (HLA), expresses class I and II molecules (alleles) that present antigens to the T-cell receptors as part of the adaptive immune response (1–4). The high level of gene sequence diversity [25,756 alleles and counting; IPD IMGT/HLA database (Release 3.38.0), http://hla.alleles.org/nomenclature/stats.html] within the HLA system may govern the outcome of many transplants (tolerance or rejection) (5, 6), infections, susceptibility to autoimmune diseases (2, 7–9), and allergic reactions to various drugs (10). Moreover, the efficacy of recently developed checkpoint inhibitory therapies in immuno-oncology appear to be directly linked to the so-called "tumor mutation burden" that is the status of neo-antigens presented by the patient's HLA class I alleles (4). During the past 20 years, although HLA allele studies have shifted from serological allele typing to molecular genotyping, most have still focused on the phenotypic description of association between diseases and HLA alleles (11, 12).

With the continuing next generation sequencing (NGS) revolution, a better understanding is slowly emerging about the diversity of the HLA genomic and transcriptomic regions including the qualitative and quantitative effects of regulatory variation on HLA expression, gene diversity, and polymorphisms (alleles) on shaping lineage-specific expression, and HLA expression on disease susceptibility and transplantation outcomes. Regulatory *cis* and *trans* polymorphisms that affect transcriptional regulation and susceptibility to complex diseases (13) are considered to be a driving force in phenotypic evolution (14, 15). Previous small-scale, low-resolution, targeted studies revealed the importance of differential allelic expression (DAE) of HLA genes in disease development and progression. Cauli et al. (16) reported a greater expression of HLA-B27 molecules in patients with ankylosing spondylitis than in healthy subjects. The association among allelic differences in HLA expression levels and disease were reported for single HLA alleles/loci such as HLA-B expression and immunoglobulin A (IgA) deficiency (17); HLA-C expression and HIV control (18–20); Crohn disease (21), and acute graft-vs.-host disease (GVHD) (22); HLA-DQ and HLA-DR expression and cystic fibrosis (23); HLA-DP expression and hepatitis B virus infection (24) and acute GVHD (25); and HLA-DRB5 and interstitial lung disease (26). In addition, suppressed or abnormal HLA expression levels were reported in gastric cancer (27), cancer cell lines (28), ovarian carcinomas (29), Merkel cell carcinoma (30), and lung cancer (31). Although polymorphisms located in the 5′ promoter region and 3′ untranslated regions (3′UTR) of HLA genes can affect HLA expression levels (21, 32–36), reliable data on HLA polymorphisms associated with HLA gene expression levels in HLA-associated disease, infection, and transplantation are still lacking.

There are different ways to measure HLA differential allele expression in leukocytes. Previously, a few particular HLA genes and alleles were examined in expression studies using flow cytometry and fluorolabeled monoclonal antibodies to measure the intensity of HLA protein surface expression (20, 21, 37) and by quantitative reverse transcription PCR (qRT-PCR) to estimate HLA transcription levels (38). Microarray methods, such as Affymetrix and Illumina, using oligoprobes are useful for the semiquantification of HLA gene transcripts expressed by a larger array of HLA class I and II genes (39, 40), but like flow cytometry and qRT-PCR, they do not identify the different HLA genotypes and alleles. In addition, all these methods are labor intensive/time consuming and often lead to ambiguous results because of problems with specificity and sensitivity and inadequate controls and reference samples. New RNA quantitative techniques based on RNA-sequencing (RNA-Seq) have emerged recently (41), and genotyping, mapping the expression quantitative trait locus, and analyzing allele-specific expression from public RNA-Seq data are promising new development (42). In addition, a computational pipeline to accurately estimate expression for HLA genes based on RNA-Seq was developed for both locus-level and allele-level estimates (43).

HLA genes also can be genotyped by amplicon sequencing using HLA transcripts as reverse-transcribed complementary DNA (cDNA) (44) and HLA RNA expression levels quantitated by amplicon sequencing using HLA locus-specific primers (45). However, the method using HLA locus-specific primers for measuring RNA levels are mostly semiquantitative because PCR efficiency can differ between the polymorphic HLA alleles (46). In contrast, a recently described capture RNA-Seq method for the quantitation of RNA expression levels of targeted genes was shown to provide enhanced coverage for sensitive gene discovery, robust transcript assembly, and accurate gene quantification (47).

In the present paper, we describe a newly developed capture RNA-Seq method for enriched NGS, genotyping, and for quantitating RNA levels of all 12 classical HLA loci [*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5* (*HLA-DRB3/DRB4/DRB5*)] and alleles using over 200 RNA samples isolated from the peripheral blood mononuclear cells (PBMCs, $n = 161$) and umbilical cord bloods (UCBs, $n = 48$) of healthy donors.

# MATERIALS AND METHODS

## Sample Information

A reference set of PBMC samples from 161 donors were selected from a larger number of high-resolution genotyped samples obtained from 2,344 donors recruited for bone marrow transplantation (BMT) as part of the Japan Marrow Donor Program (JMDP) (48). This sample of genotyped Japanese donors represented more than 99.2% cumulative allele frequency (138 alleles) of the known HLA alleles in Japanese population at the field-2 level of resolution (an allele resolution level of sequences that differs by a non-synonymous substitution) at six HLA loci (20 alleles for *HLA-A*, 37 *HLA-B*, 19 *HLA-C*, 30 *HLA-DRB1*, 16 *HLA-DQB1*, and 16 *HLA-DPB1*) (**Table S1**). A total of 48 UCB samples previously registered and stored at the Tokai University Cord Blood Bank also were used in this study. The high-resolution HLA genotyping data for these UCB was unknown before this study. The mononuclear cells of the UCB were isolated by Ficoll–Paque density separation (Ficoll–Paque$^{TM}$ Plus, GE Healthcare).

## Isolation of RNA Samples and Measurement

Total RNA was isolated from the PBMC and UCB mononuclear cells using TRIzol (Thermo Fisher Scientific). The quantity and quality of the RNA were determined using an RNA 6000 Nano Kit with a Bioanalyzer 2100 (Agilent Technologies).

## Design of Sequence Capture Probes

The customized biotinylated nucleotide capture probes were designed and synthesized by Roche's proprietary method for use with the SeqCap RNA Enrichment System (Roche, NimbleGen, KAPA Biosystems). The exact number of 5′-biotinylated probes used per gene locus for this study was not released to us by the manufacturer. However, each of the single-stranded, 5′-biotinylated capture oligonucleotide probes in the synthesized set was 50–100 bases (average 75 bases) in length. Taken together, all of the capture probes in the set represented sequences of 172 alleles (19 *HLA-A*, 39 *HLA-B*, 19, *HLA-C*, 2 *HLA-DRA*, 31 *HLA-DRB1*, 3 *HLA-DRB3*, 3 *HLA-DRB4*, 3 *HLA-DRB5*, 16 *HLA-DQA1*, 15 *HLA-DQB1*, 4 *HLA-DPA1*, and 18 *HLA-DPB1*) that were representative of the Japanese population (**Table S2**) (49). Of the 172 allelic targets, 160 covered full-length HLA regions and 12 sequences covered partial regions such as specific exons and/or intron regions, respectively. The total nucleotide length covered by the designed probes was 1,321,811 bp, which covered 96.1% (1,270,384 bp) of the targeted regions. The remaining 3.9% of the targeted regions were omitted from the probe design and synthesis because of repeat sequences.

## Sequence Capture and Next-Generation Sequencing

The workflow for the Capture RNA-Seq profiling by NGS is shown in **Figure S1**. The basic steps were (1) RNA fragmentation, (2) preparation of reverse-transcribed RNA libraries, (3) hybridization with biotin-labeled capture probes containing target sequences, (4) capture and enrichment of targeted sequences with streptavidin-coated paramagnetic beads, (5) library amplification, (6) NGS Illumina sequencing, and (7) data analysis for HLA allele assignments and quantitation of allelic sequence expression (**Figure S1A**).

Total RNA (100 ng) was fragmented by shearing and reverse transcribing into cDNAs by second-strand synthesis using a KAPA Stranded RNA-Seq Library Preparation Kit. The sheared product was purified by Agencourt AMPure XP reagent (Beckman Coulter), and the cDNA libraries were constructed using KAPA library preparation kits (KAPA Biosystems) with SeqCap Adapter Kits A and B (Roche Life Science). The cDNA libraries were sized and quantitated using an Agilent DNA 1000 Kit with the Bioanalyzer 2100 (Agilent Technologies). One hundred nanograms of each indexed library was pooled together according to the manufacturer's recommendation. The pooled library was mixed with a SeqCap HE universal oligonucleotide, SeqCap HE-Oligo Kits A and B (Roche Life Science), and COT-1 human DNA and then vacuum evaporated at 60°C for ∼30 min. The custom-designed sequence capture probe set was added and hybridized at 47°C for 18 h using a SeqCap Hybridization solution in a GeneAmp PCR System 9700 (Thermo Fisher Scientific). After hybridization, the non-specific hybridization products were washed out with a Wash Kit, and the captured library was enriched with a SeqCap Pure Capture Bead Kit (Roche/NimbleGen). The enriched beads were subjected directly to post-capture amplification by ligation mediated (LM) PCR using a SeqCap EZ accessory kit v2 (Roche/NimbleGen). The enriched and amplified NGS library was purified by AMPure XP reagent, quantitated using the Agilent DNA 1000 Kit with the Bioanalyzer 2100, and sequenced using MiSeq Reagent Kit v.2 (300 cycles), generating 150 bp pair-end sequence reads with an Illumina MiSeq System according to the manufacturer's protocol (Illumina).

## Data Processing and Allele Assignment of HLA Genotypes

After the sequencing runs, basic sequence information such as the read numbers and quality values were calculated with a FASTX_quality_stats program included in a FASTX-Toolkit package (ver. 0.0.13) for short-read data preprocessing (http://hannonlab.cshl.edu/fastx_toolkit/).

The FASTQ sequence files for each sample were used for HLA genotyping and allele assignment up to the field-3 level (an allele resolution where synonymous and/or non-synonymous DNA substitutions in the coding region define alleles). The HLA alleles for the 12 classical HLA loci, *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB/DRB4/DRB5*, *HLA-DQB1*, *HLA-DQA1*, *HLA-DPA1*, and *HLA-DPB1*, were assigned by nucleotide similarity searches in the IPD-IMGT-HLA database (http://hla.alleles.org/) using the BLAT program (50), included in an in-house Sequence Alignment Based Assigning Software (SeaBass) (51). When novel single-nucleotide polymorphisms (SNPs) were detected, they were confirmed by Sanger direct sequencing using newly designed sequencing primers.

## Calculation and Normalization of Sequence Read Numbers (RNA Levels)

Mapping of the reads and the HLA allele sequences assigned by the BLAT search as references were performed using GS Reference Mapper Ver. 3.0 software (Roche). To precisely extract in-phase read numbers that are our measure of RNA levels, we limited the mapping regions to highly polymorphic exons: 546 bp of exons 2 and 3 in *HLA-A*, *HLA-B*, and *HLA-C*; 239 bp of exon 2 in *HLA-DRA*; 270 bp of exon 2 in *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, and *HLA-DQB1*; 249 bp of exon 2 in *HLA-DQA1*; 246 bp of exon 2 in *HLA-DPA1*; and 264 bp of exon 2 in *HLA-DPB1*. The mapping parameter was set to a 100% matched condition between the reads and the references to avoid mismapping among the HLA loci and contamination of *in vitro* generated PCR crossover products (51, 52).

In order to compare differences of the RNA levels (cDNA sequence read numbers) among the HLA loci and alleles, we normalized the mapped read numbers as follows. The mapped raw read numbers of each allele at each locus were first standardized by their target sizes. To derive the "normalized read numbers" within a particular set of loci, the total size-standardized read numbers included in the set were estimated, and the total read numbers were standardized to 1 million. Then, the normalized read numbers of each allele at each locus in the set were calculated as relative read numbers in 1 million total reads. An example of deriving the normalized read numbers at the *HLA-A*, *HLA-B*, and *HLA-C* loci is described in **Figure S1B**.

For comparison of the RNA levels expressed by the same or different HLA genes and alleles, we prepared the following four datasets: (1) normalized read numbers of the 12 HLA loci and alleles (including two *HLA-DRB3/DRB4/DRB5* alleles) using 78 PBMC and 18 UCB samples, (2) normalized read numbers of the HLA class I loci and alleles using 161 PBMC and 48 UCB samples, (3) normalized read numbers of the HLA class II loci (excluding *HLA-DRB3/DRB4/DRB5*) and alleles using 161 PBMC and 48 UCB samples, and (4) normalized read numbers of the HLA class II loci and alleles (including two *HLA-DRB3/DRB4/DRB5* alleles) using 78 PBMC and 18 UCB samples (**Figure S1C**).

Comparative analyses of the read numbers (RNA levels) among the HLA loci and among the HLA alleles were carried out using only heterozygous alleles that were completely phased in each dataset (**Figure S1C**). We excluded homozygous alleles, partially phased alleles, and hemizygous alleles from this study to avoid mapping biases because it was not possible to divide these reads into unphased exonic regions.

## Statistical Analyses

The results of the comparative analyses of capture RNA sequences were drawn as a box-and-whisker diagram using the graphics output of Microsoft Excel 2016. The box-and-whisker diagram displays the median with upper and lower quartiles within the box, with the whiskers extended 1.5 times the interquartile range from the box. Statistical differences of the RNA levels expressed by the HLA loci and alleles were calculated by analysis of variance (ANOVA) in Microsoft Excel 2016. A two-sided $p < 0.05$ was considered statistically significant. Bias in statistical significance due to multiple calculations was not taken into account. Correlation coefficients, coefficient of determination, and approximate curve between different sets of results were calculated by the Pearson correlation coefficient method using the Microsoft Excel function (Excel for Mac version 16.25).

# RESULTS

## Assay Design

We devised a capture-based RNA-Seq NGS method to simultaneously analyze DNA alleles and RNA expression levels of HLA genes as described in section Material and Methods. Because target sequences have a high degree of polymorphisms, we designed a capture probe set, which covers 172 most frequent Japanese HLA allele sequences. The NGS data that we obtained were processed by an in-house program, in which only the reads showing perfect matches (100% identity) to reference sequences were included in the analyses. The relative levels of RNA expression of each HLA gene were deduced from read numbers, which were normalized against total read numbers of a set of HLA genes, among which RNA expression levels were compared (see section Material and Methods). Thus, we obtained four data sets: set 1, composed of all 12 HLA loci; set 2, composed of *HLA-A*, *HLA-B*, and *HLA-C* loci; set 3, composed of class II loci excluding *HLA-DRB3/DRB4/DRB5* genes; set 4, composed of class II loci including *HLA-DRB3/DRB4/DRB5* genes (see section Material and Methods; **Figures S1B,C**).

## Sequence Read Information for PBMC and UCB Samples

Sequence read information was obtained for all the captured and enriched sequence libraries constructed from 161 PBMC and 48 UCB. In PBMCs, the total draft read numbers were 281,256,904 reads with a range of reads from 398,390 to 3,196,406 reads [1,746,937 ± 457,246 standard deviation (SD) on average]. These were high-quality reads with quality values (QV) of >30 with an average QV of 35.7 ± 0.5. The draft read bases were 38.1 Gb in total with a range between 55.8 and 428.1 Mb (237.0 ± 62.6 Mb on average) and with an overall average read length of 135.6 ± 4.4 bases. In contrast, in the UCB, the total draft read numbers were 82,231,970 reads ranging from 813,198 to 7,294,988 (1,713,166 ± 900,625 on average) of high-quality reads with QV > 30 and an average QV of 36.4 ± 0.1. The total draft read bases were 11.2 Gb with a range between 111.2 and 1,007.9 Mb (232.3 ± 124.6 Mb on average) and an overall average read length of 135.5 ± 3.9 bases. Therefore, these high-quality reads had sufficient sequencing volume for further HLA genotyping analysis.

## Genotyping to the Field-3 Level for 12 HLA Loci

Nucleotide similarity searches of the sequenced HLA alleles to the field-3 level (based on synonymous and/or non-synonymous substitution in the coding region for each designated allele) using the BLAT program (50) identified a total of 177 alleles for the 161 PBMC, including 21 *HLA-A*, 39 *HLA-B*, 19 *HLA-C*, 4 *HLA-DPA1*, 16 *HLA-DPB1*, 18 *HLA-DQA1*, 16 *HLA-DQB1*, 2 *HLA-DRA*, 30 *HLA-DRB1*, 5 *HLA-DRB3*, 3 *HLA-DRB4*, and 3 *HLA-DRB5*; and 114 alleles for the 48 UCB samples including 12 *HLA-A*, 22 *HLA-B*, 14 *HLA-C*, 3 *HLA-DPA1*, 10 *HLA-DPB1*, 11 *HLA-DQA1*, 13 *HLA-DQB1*, 2 *HLA-DRA*, 18 *HLA-DRB1*, 4 *HLA-DRB3*, 3 *HLA-DRB4*, and 2 *HLA-DRB5* alleles with no phase ambiguity (**Table S3**). One novel allele was further detected for *HLA-DQA1* [named tentatively *DQB1*06:02new* (DDBJ Accession Number: LC499658) with a synonymous substitution in the exon 1]. In contrast, the other alleles were previously known HLA alleles that were assigned to the field-2 level at six HLA loci.

## Comparison of the RNA Levels for 12 HLA Loci and Alleles in PBMC Samples

**Figure 1** shows a box-and-whisker diagram of the normalized read numbers (RNA levels) for 12 HLA loci using dataset 1 that includes 1,252 phased alleles identified for the PBMC samples (**Table S4A**). The ANOVA statistical difference between the expression levels of the HLA loci was highly significant

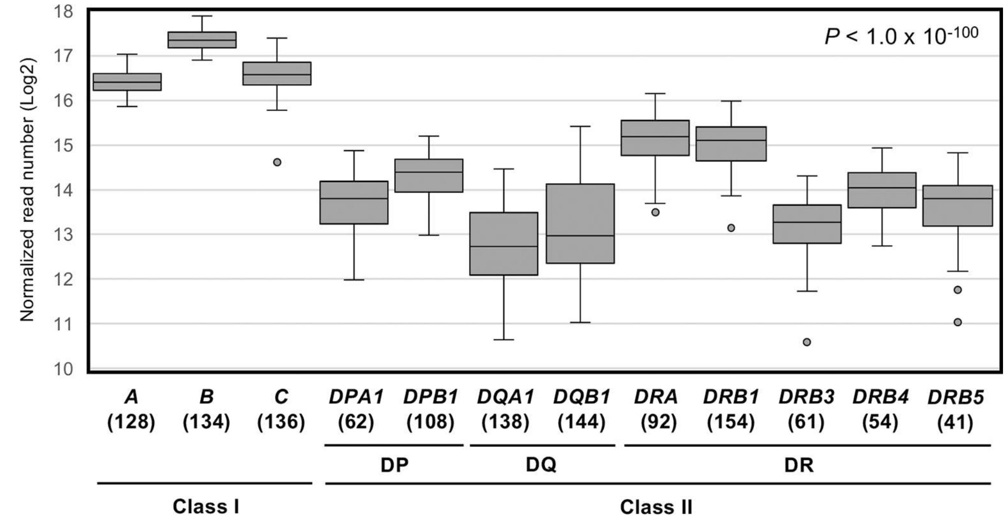


**FIGURE 1 |** RNA expression levels of 12 human leukocyte antigen (HLA) loci in PBMC samples measured by capture RNA-Seq. These box-and-whisker diagrams were drawn using the dataset 1 obtained from the sequence reads of 78 PBMC samples (**Figure S1C**). Vertical axis indicates normalized read numbers (Log2) calculated as described in section Material and Methods. Horizontal axis indicates the 12 classical class I and class II HLA loci. Horizontal lines in the boxes indicate median level of expression at each locus. Parenthesis below the locus name indicates the number of individual points plotted per each locus.

**TABLE 1 |** Five-number summary of normalized RNA levels at each human leukocyte antigen (HLA) locus in peripheral blood mononuclear cell (PBMC).

| Class | Subregion | Locus | Minimum reads | First quartile reads | Median reads | Third quartile reads | Maximum reads | Ratio of hinges* |
|---|---|---|---|---|---|---|---|---|
| Class I | | *HLA-A* | 59,565 | 75,996 | 86,788 | 100,109 | 134,318 | 1.3 |
| | | *HLA-B* | 121,917 | 148,807 | 165,092 | 187,736 | 242,948 | 1.3 |
| | | *HLA-C* | 56391 | 84,675 | 98,672 | 120,247 | 172,204 | 1.4 |
| Class II | DP | *HLA-DPA1* | 4,032 | 8,058 | 12,585 | 15,217 | 24,991 | 1.9 |
| | | *HLA-DPB1* | 8,101 | 15,232 | 20,354 | 25,442 | 34,833 | 1.7 |
| | DQ | *HLA-DQA1* | 1,602 | 4,623 | 7,744 | 12,744 | 22,625 | 2.8 |
| | | *HLA-DQB1* | 2,083 | 5,201 | 8,096 | 18,605 | 43,870 | 3.6 |
| | DR | *HLA-DRA* | 12,069 | 27,862 | 36,511 | 46,162 | 69,618 | 1.7 |
| | | *HLA-DRB1* | 9,046 | 25,641 | 35,606 | 43,578 | 64,661 | 1.7 |
| | | *HLA-DRB3* | 1,535 | 7,164 | 9,972 | 13,113 | 20,243 | 1.8 |
| | | *HLA-DRB4* | 6,819 | 12,271 | 15,445 | 21,270 | 31,237 | 1.7 |
| | | *HLA-DRB5* | 2,101 | 9,093 | 13,634 | 17,941 | 29,122 | 2.0 |

**Hinges mean first and third quartiles. Ratio of hinges was calculated by third quartile reads/first quartile reads.*

($P < 1.0 \times 10^{-100}$). We also found that the 99% confidence intervals for these 12 loci do not overlap (data not shown), confirming that each 12 loci are distinct in terms of RNA expression levels.

Among the 12 loci compared here, *HLA-B* displayed the highest average expression level. The average level of expression of *HLA-B* was ∼2-fold higher than that of *HLA-A* or *HLA-C*, and ∼4–5 times higher than by the class II loci, *HLA-DRA*, *HLA-DRB1*, or others (**Table 1**). In addition, the average expression levels of *HLA-DRB4* and *HLA-DRB5* were 1.4–1.5-fold higher than *HLA-DRB3* (**Figure 1** and **Table 1**). The lowest reads were for *HLA-DQA1* and *HLA-DQB1*. The ratio of hinges (third quartile reads/first quartile reads) in the box-and-whisker diagram confirmed that locus-specific variations of the read numbers for the *HLA-DQ* RNA levels (2.8 for *HLA-DQA1* and 3.6 for *HLA-DQB1*) were much higher than 1.3–1.4 for the class I RNA levels and 1.7–2.0 for the *HLA-DP* and *HLA-DR* RNA levels (**Table 1**).

The RNA levels expressed at each allele were analyzed using the read numbers obtained from at least three different samples. **Figure 2** shows box-and-whisker diagrams of the RNA levels for the HLA alleles of PBMC samples using datasets 2–4 that include 2,175 (763 class I + 1,260 class II + 152 *HLA-DRB3/DRB4/DRB5*) heterozygous alleles (**Table S4A**). The DAE was observed for all HLA class I and II genes, except for *HLA-DPA1* and *HLA-DRA*, with strong statistical significant differences (ANOVA) for RNA levels expressed by the HLA alleles of *HLA-B* ($P = 2.1 \times 10^{-15}$) to *HLA-DQB1* ($P = 5.1 \times 10^{-95}$) (**Figures 2A–J**). In addition, the ratios between the lowest and highest expressed alleles at each locus showed that *HLA-DQA1* and *HLA-DQB1* had the largest allelic differences with 3.8 ($P = 5.0 \times 10^{-11}$) and 5.8 ($P = 1.4 \times 10^{-15}$), respectively, and that there were no significant allelic differences for *HLA-DPA1* and *HLA-DRA* (**Figures 2D–J** and **Table 2**). Consistently, when degrees of DAE are compared for coefficient of variation (CV; standard deviation divided by mean), the class II *HLA-DQA1* and *HLA-DQB1* loci displayed highest levels of allelic differences compared to others, including all the class I loci, as shown in **Figure S2**.

Based on the normalized read counts, we next compared variations among samples with the same alleles. As depicted in **Figure S2**, CV of read counts obtained from the samples with the same alleles at class I are relatively lower than those at class II, as the three class I loci showed the three lowest averaged CV values among all the 12 loci. It appears, therefore, that the class I genes are more evenly expressed among individuals with the same alleles, whereas class II shows more variations in expression levels among individuals with the same alleles. The larger extents of intra-allelic variations observed with the class II genes could be due to different distribution patterns of subcell populations, expressing the class II genes differently, among the samples with the same alleles.

## HLA Polymorphisms and RNA Expression Levels in Specific HLA Loci and Haplotypes of PBMC Samples

### Relationship Between HLA-DR Haplotypes and RNA Levels

Since the correlation between *HLA-DRB1-HLA-DRB3/DRB4/DRB5* haplotypes and RNA expression levels was unknown, we examined 31 *DRB3/DRB4*, 18 *DRB3/DRB5*, and 23 *DRB4/DRB5* heterozygous samples (a total of 144 *DRB1-DRB3/DRB4/DRB5* haplotypes) from dataset 4 (**Table S4A**). In this dataset, there was a total of 28 *DRB1-DRB3/DRB4/DRB5* haplotypes with 14 assigned as *DRB1-DRB3*, 11 as *DRB1-DRB4*, and 3 as *DRB1-DRB5*. These haplotypes were identified by estimating *HLA-DRB1* and *HLA-DRB3/DRB4/DRB5* alleles without observing any discrepancies to previously reported HLA-DR genomic structures (53, 54).

**Figure S3A** shows the comparative relationships of RNA levels expressed by the *HLA-DRB1* and *HLA-DRB3/DRB4/DRB5* loci using 17 *DRB1-DRB3/DRB4/DRB5* haplotypes that were analyzed using at least three different samples (**Table S5**). The RNA levels expressed by *HLA-DRB1* were widely distributed for all haplotypes, whereas the expression levels of *HLA-DRB4* and *HLA-DRB5* tended to be higher than for *HLA-DRB3* (**Figure S3A**). In addition, there was no significant correlation between the RNA expression levels of *HLA-DRB1* and *HLA-DRB3/DRB4/DRB5* ($R^2 = 0.0003$), indicating that they are regulated independently of each other in PBMC samples.

For RNA levels expressed by the *DRB1-DRB3/DRB4/DRB5* haplotypes (**Table S5**), there was a significant difference ($P = 0.0119$) between the median read numbers of *DRB1*14:06:01/DRB3*02:02:01* (median, 36,822) and *DRB1*14:54:01/DRB3*02:02:01* (median, 18,330). There was a significant difference ($P = 0.0182$) between *DRB1*09:01:02* haplotypically linked to either *DRB4*01:03:01* or *DRB4*01:03:02* (**Table S5**), but no significant difference ($P = 0.9089$) between the haplotypes *DRB1*09:01:02/ DRB4*01:03:01* and *HLA-DRB1*04* or *HLA-DRB1*07* linked to *DRB4*01:03:01*. These data show that the variance of RNA expression levels of haplotypes detected at the allelic field-2 level such as *DRB1*09:01/DRB4*01:03* can be differentiated significantly at the allele field-3 level such as *DRB1*09:01:02/ DRB4*01:03:01* or *DRB1*09:01:02/ DRB4*01:03:01*.

### Relationships Between HLA-DQ Haplotypes and RNA Expression Levels

Since significant differences were observed for RNA levels expressed by the *HLA-DQA1* and *HLA-DQB1* alleles (**Figure 2** and **Table 2**), we investigated the correlation between HLA-DQ haplotypes (*DQA1-DQB1*) and RNA expression levels. The distribution in the level of expression based on read numbers for the *HLA-DQA1* alleles ranged between low expression for *DQA1*01/05* alleles, intermediate expression for *DQA1*02/04/06* alleles, and high expression for *DQA1*03* alleles. The DAE for *HLA-DQB1* ranged from low expression for *DQB1*02/03/04* to high expression for *DQB1*05/06*. There were significant differences ($P < 0.001$) between most allelic groups in the low, intermediate, or high levels of expression (**Figure S3B**). These data were consistent with a previous report on the differential expression of *HLA-DQ* alleles in PBMC where the alleles were associated with susceptibility to and protection from type 1 diabetes (55). Interestingly, although the *DQA1-DQB1* haplotypes were composed of the highest expression group of *HLA-DQA1* alleles and the lowest expression group of *HLA-DQB1* alleles, these haplotypes occur at the highest frequency (40.5%) in the Japanese population (**Figure S3B**). In contrast, *DQA1-DQB1* haplotypes that are composed of highest expression groups of *HLA-DQA1* and *HLA-DQB1* alleles were not observed in the Japanese population. This finding suggests a trend for selection to low and at best intermediate expression of the DQ heterodimers. It is indeed intriguing that the highest expression groups were not observed in worldwide populations (17th IHIW: NGS HLA genotyping data, http://17ihiw.org/17th-ihiw-ngs-hla-data/) (**Figure S3B**).

### RNA Expression Levels of *HLA-DPB1* Alleles and Genotypes of rs9277534

The SNP marker rs9277534 is located within the 3′UTR of *HLA-DPB1*, and the RNA levels expressed by the AA genotype were reportedly significantly ($P < 0.001$) lower than those expressed by the GG genotype (**Figure S3C**) (25). In our analysis of the same genotypes by the capture RNA-Seq method, the read numbers for the AA genotype using 12 samples and the GG genotype using 24 samples that were selected from dataset 3 of PBMC samples (**Table S4A**) produced a box-and-whisker diagram (**Figure S3C**) that was similar to the results of the previous report (25, 56) with a significant difference ($P = 0.0019$) of expression between the AA and GG genotypes. However, this result was not consistent with individual alleles shown in **Figure 2E** because *DPB1*04:01:01* (low) and *DPB1*05:01:01* (high) are thought to be outliers, while *DPB1*04:02:01*, *DPB1*02:01:02* and *DPB1*02:02* that are supposed to be low expression levels have similar

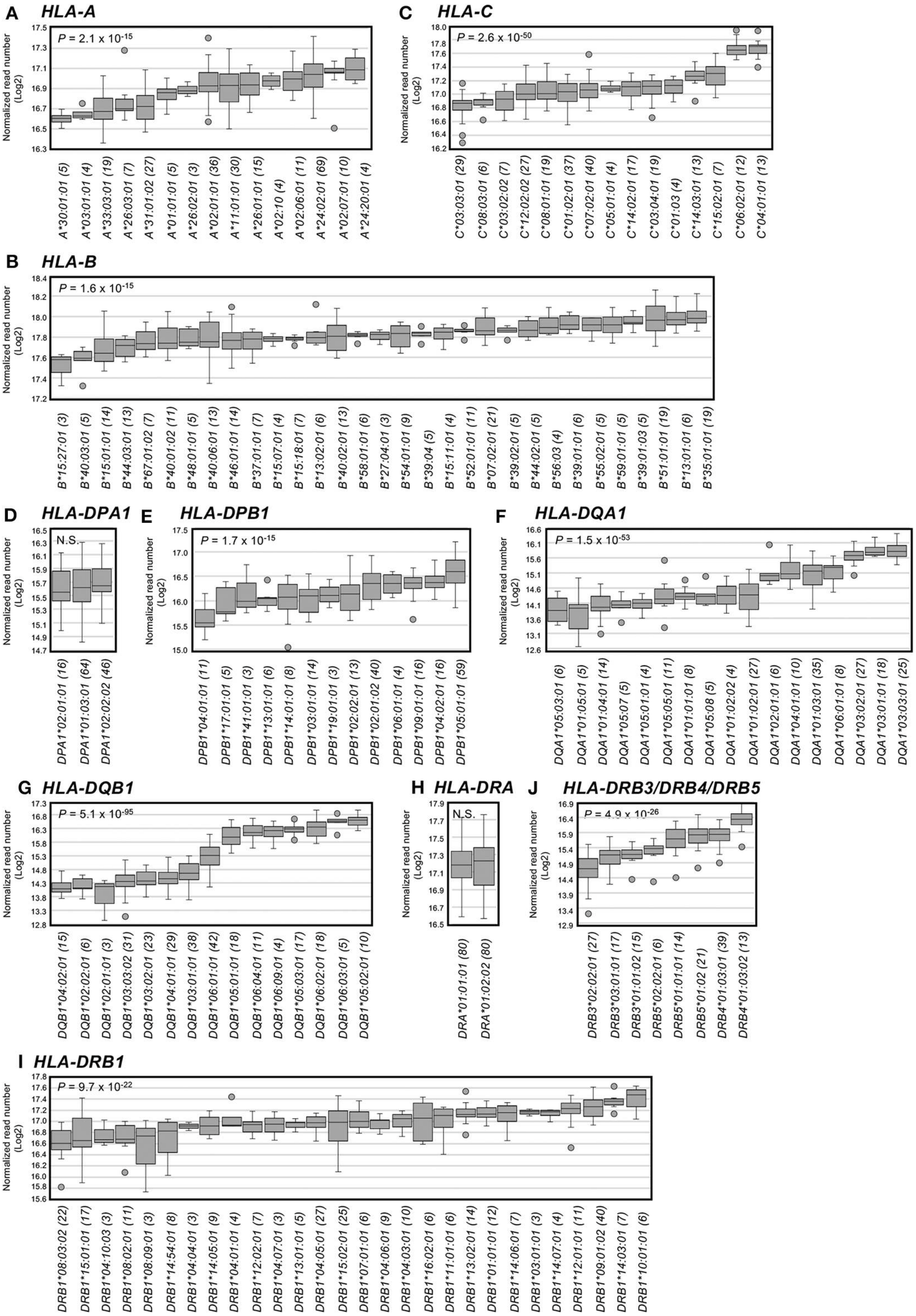


**FIGURE 2 |** Allelic RNA levels expressed by 12 human leukocyte antigen (HLA) loci in peripheral blood mononuclear cell (PBMC) samples and measured by capture RNA-Seq. The box-and-whisker diagrams were drawn using the datasets 2–4 obtained from the sequence reads of 78 or 161 PBMC samples (**Figure S1C**). Panels **(A–J)** show **(A)** *HLA-A*, **(B)** *HLA-B*, **(C)** *HLA-C*, **(D)** *HLA-DPA1*, **(E)** *HLA-DPB1*, **(F)** *HLA-DQA1*, **(G)** *HLA-DQB1*, **(H)** *HLA-DRA*, **(I)** *HLA-DRB3/DRB4/DRB5*, and **(J)** *HLA-DRB1*. Vertical axis indicates normalized read numbers (log2) calculated as described in section Material and Methods. Horizontal axis indicates the alleles for each of HLA loci **(A–J)**. Horizontal lines in the box indicate median expression at each allele. Parenthesis following the allele name indicates the number of individual points plotted per each locus.

**TABLE 2 |** Comparison of the lowest and highest expressed alleles of each locus using normalized reads at median in peripheral blood mononuclear cells (PBMCs).

| Locus | Average median reads | The lowest expressed allele | | The highest expressed allele | | Fold change* | *P*-value |
|---|---|---|---|---|---|---|---|
| | | Allele | Median reads | Allele | Median reads | | |
| *HLA-A* | 120,366 | *A*30:01:01* | 99,814 | *A*24:20:01* | 139,413 | 1.4 | $6.7 \times 10^{-4}$ |
| *HLA-B* | 231,598 | *B*15:27:01* | 196,005 | *B*35:01:01* | 258,664 | 1.3 | $2.3 \times 10^{-6}$ |
| *HLA-C* | 142,987 | *C*03:03:01* | 119,834 | *C*04:01:01* | 212,951 | 1.8 | $2.1 \times 10^{-18}$ |
| *HLA-DPA1* | 49,804 | *DPA1*02:01:01* | 48,111 | *DPA1*02:02:02* | 51,134 | 1.1 | NS |
| *HLA-DPB1* | 73,344 | *DPB1*04:01:01* | 47,828 | *DPB1*05:01:01* | 99,193 | 2.1 | $1.3 \times 10^{-10}$ |
| *HLA-DQA1* | 29,024 | *DQA1*05:03:01* | 14,991 | *DQA1*03:03:01* | 57,501 | 3.8 | $5.0 \times 10^{-11}$ |
| *HLA-DQB1* | 50,644 | *DQB1*04:02:01* | 17,371 | *DQB1*05:02:01* | 99,929 | 5.8 | $1.4 \times 10^{-15}$ |
| *HLA-DRA* | 151,031 | *DRA*01:01:01* | 148,601 | *DRA*01:02:02* | 153,460 | 1.0 | NS |
| *HLA-DRB1* | 132,803 | *DRB1*08:03:02* | 99,464 | *DRB1*10:01:01* | 182,883 | 1.8 | $6.3 \times 10^{-9}$ |
| *HLA-DRB3* | 33,383 | *DRB3*02:02:01* | 26,715 | *DRB3*01:01:02* | 37,037 | 1.4 | $4.3 \times 10^{-3}$ |
| *HLA-DRB4* | 70,578 | *DRB4*01:03:01* | 58,794 | *DRB4*01:03:02* | 82,362 | 1.4 | $2.5 \times 10^{-2}$ |
| *HLA-DRB5* | 50,935 | *DRB5*02:02:01* | 42,508 | *DRB5*01:02* | 57,684 | 1.4 | $6.0 \times 10^{-6}$ |

**Fold change was calculated by reads of the highest expressed allele/reads of the lowest expressed allele.*

RNA levels as those of *DPB1*09:01:01* and *DPB1*06:01:01* that are supposed to be high expression levels (57). In the case of excluding the outlier alleles, the significant difference was not obtained between the AA and GG genotypes (**Figure S3C**). These data suggested that not only this SNP was associated with HLA-DP expression levels but also the other (one or more SNPs) can be involved in the HLA-DP expression levels.

### RNA Expression Levels of Null or Low Expressed HLA Alleles

Of over 25,000 HLA alleles released from the IPD-IMGT/HLA database (http://hla.alleles.org/), ∼1,000 alleles are characterized as null and low expressed alleles. We tested if the RNA-Seq method is able to discriminate those null alleles by analyzing read data corresponding to the *HLA-A*02:15N* and *A*02:53N* alleles, which have been categorized as null. Examination of the RNA levels expressed by the known *HLA-A* null alleles *A*02:15N* and *A*02:53N* (**Figure S3D**) revealed that they were at 34.3% (41,242 reads) and 18.2% (21,902 reads), respectively, of the average median of *HLA-A* allelic expression (120,366 reads in **Table 2**). This result suggests that the capture RNA-Seq can directly identify null and low expression allele in a quantitative manner.

## Comparison of the HLA RNA Levels in PBMC and UCB

A box-and-whisker diagram of the normalized read numbers (RNA levels) expressed by 12 HLA loci (**Figure S4A**) in UCB was constructed using the dataset 1 that includes 280 phased alleles (**Table S4B**). The ANOVA statistical difference between the expression levels of the 12 HLA loci of UCB was highly significant (ANOVA, $P < 1.0 \times 10^{-95}$) similarly to those of PBMC (**Figure 1** and **Figure S4A**). The DAE was observed for all HLA class I and II genes, except for *HLA-DRA* and *HLA-DRB5*. Collectively weak statistical significances were also noted for RNA levels expressed by the HLA alleles of *HLA-DPA1* ($P = 1.2 \times 10^{-2}$) to *HLA-DQB1* ($P = 1.6 \times 10^{-31}$) (**Table S6**). The observation of higher *p* values in UCB than in PBMC is thought to depend on the number of samples. To find differences in the RNA expression levels between the HLA alleles of PBMC and UCB, we compared 30 class I and 44 class II alleles using the normalized reads of 590 alleles (206 class I and 384 class II alleles) in UCB (**Table S4B**) and the corresponding 1,724 alleles (583 class I and 1,141 class II alleles) in PBMC (**Table S7**). **Figure 3** shows a high correlation for the RNA levels expressed by HLA alleles of PBMC and UCB. The RNA levels expressed by the PBMC and UCB of HLA class I and II loci were correlated strongly with a coefficient of determination of $R^2 = 0.9319$ and 0.9486, respectively (**Figures 3A,B**). However, there were significant differences between PBMC and UCB for HLA RNA levels expressed by 20 of the 74 alleles (**Table S7**). Most significant differences ($P < 0.05$) occurred for *HLA-A* (2 of 8 alleles), *HLA-B* (4 of 13 alleles), *HLA-C* (5 of 9 alleles), and *HLA-DPB1* (4 of 5 alleles) with no differences for 9 alleles of *HLA-DQA1* and only a few differences for the alleles of the other HLA loci. Comparatively large differences ($P < 0.001$) of RNA levels between the PBMC and UCB were observed for seven HLA alleles, *A*11:01:01*, *C*07:02:01*, *DPA1*01:03:01*, *DPB1*05:01:01*, *DQB1*04:02:01*, *DRB1*04:05:01*, and *DRB1*01:01:01* (**Figure S4B** and **Table S7**). The results indicated that HLA expression patterns in PMBC and UCB are similar but not entirely the same with some differences. The differences are more pronounced in particular alleles, suggesting that those alleles are regulated by separated factors (genes) between PMBC and UCB.

## Validation of the Capture RNA-Seq Method

In an evaluation of our capture RNA-Seq method, **Figure S5** shows the high correlation ($R^2 = 0.9135$) that we obtained when we compared our quantitative HLA gene expression results to those previously obtained by Fagerberg et al. (58) using a RNA-Seq method to quantitate HLA gene expression levels in white blood cells (WBC) for a Human Body Map project (NCBI BioProject Accession: PRJEB2445). This comparison between different expression methods confirmed the reliability

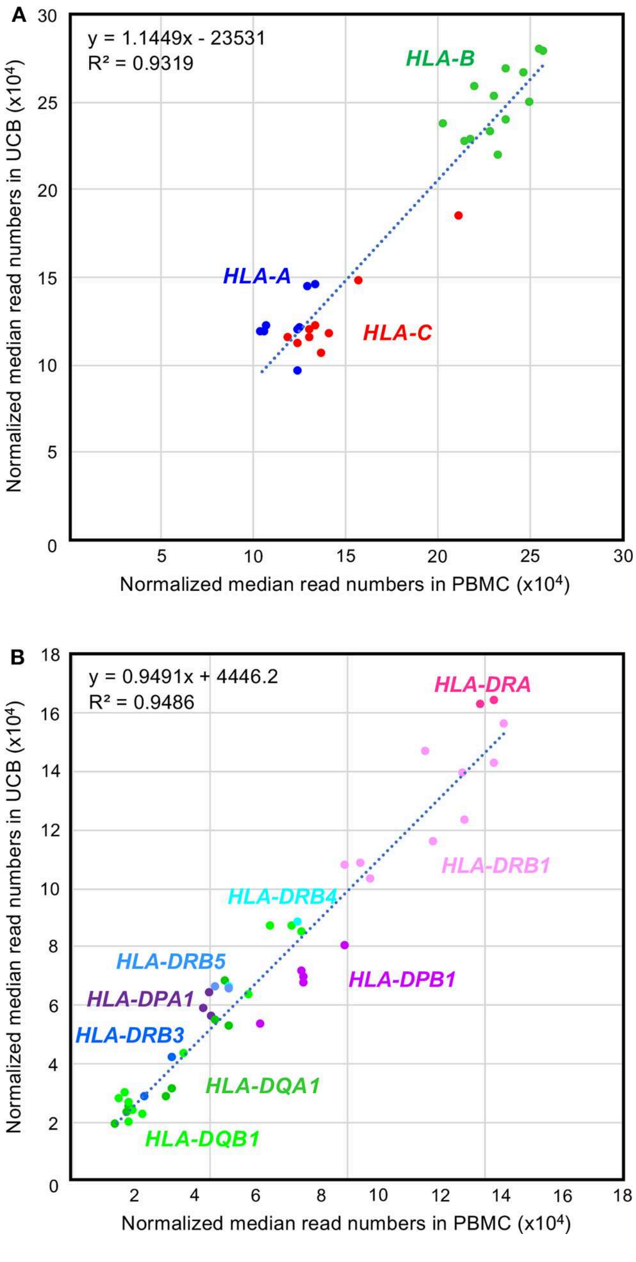


**FIGURE 3 |** Comparison between the RNA expression levels for peripheral blood mononuclear cell (PBMC) and umbilical cord blood (UCB) samples. **(A,B)** Correlations of the RNA expression levels in the class I and class II loci, respectively. Normalized median reads (×$10^4$) in PBMC and UCB (**Table S7**) were plotted in the horizontal and vertical axis, respectively.

and accuracy of quantitating HLA gene expression using our capture RNA-Seq method.

We also tested to assess whether the allelic read–number differences might be generated by methodological artifacts, such as differences in probe capturing efficiencies among different alleles, by estimating allelic read-number ratios of the original quantitation targets for individual samples as well as the allelic read-number ratios of subregions of the original targets. The results, displayed in **Figure S6** and showing comparisons of the allelic read number ratios of the original targets and those of their subregions for individual samples, indicated a good correlation between ratios for the original targets and those for their subregions. We, therefore, concluded that the allelic differences in the read numbers were not generated by artifacts and most likely reflect allelic differences in RNA levels. We, however, noted that small allelic differences, such as those observed with the HLA-B genes, could be derived in part from artifacts such as allelic differences in capturing efficiencies or from experimental errors.

## DISCUSSION

In this study, we developed a protocol for genotyping 12 classical HLA genes and quantitating their gene and/or allelic expression using a single-capture RNA-Seq method and demonstrated the usefulness of the method in a comparative analysis of the RNA expression levels among the 12 HLA loci and 2,850 alleles using 161 PBMC and 48 UCB samples. For the RNA expression analysis, we chose to use previously HLA genotyped PBMC samples that had been collected as part of the Japan Marrow Donor Program (48) from medically well-tested BMT donors prior to transplantation. Therefore, these HLA-genotyped PBMC samples were from well-characterized healthy donors who were considered to be free from infection, cancer, inflammatory diseases, and other ailments that could have changed the normal baseline levels of HLA gene expression. In order to test our RNA sequencing method before applying it to disease and transplantation research, it was considered important to first understand comprehensively the normal pattern of RNA levels expressed by each allele using healthy individuals within the same ethnic group. In this way, the differences in RNA expression levels (sequence read numbers) could be compared more reliably among the HLA alleles at the different HLA loci as described previously (43).

An important aspect of our RNA-Seq method was the design and application of 172 capture RNA probe sequences of 172 most frequent Japanese HLA allele sequences that would be used to capture and enrich the targeted HLA alleles in the blood sample. If we had used only single representative allelic probes (reference probes) for each HLA gene, such as those represented by the latest version of human genome reference (e.g., GRCh38.p12 at NCBI), the target DNA fragments might not be enriched or might have been missed due to low affinity between the reference-derived RNA probe and the target HLA allele-derived DNA fragment. As a result of poor hybridization between probe and a targeted allele, the RNA expression level of a particular allele might be recorded incorrectly to be low when in fact its allelic expression was intermediate or high. Therefore, in order to minimize the false calls of a low expression, we designed the sequences of the RNA capture probes based on our previously determined HLA allele sequences in the Japanese population (49). In addition, since 93% of the 172 allele sequences cover the full length of HLA genes (5′ promoter region to 3′UTR), the 172 RNA capture probes are useful for identifying the gene sequences of disease-specific splicing variants and gene expression-related variants and for detecting antisense RNAs such as microRNAs (miRNAs) that suppress gene expression (17).

Although specific RNA capture probes are an important first step for accurate and reliable quantitation of allele expression by the capture RNA-Seq method, the NGS and analytical bioinformatic methods for obtaining correct read numbers for RNA (cDNA) sequences expressed by particular alleles are equally important steps in the overall accuracy and reliability of the protocol. In general, it is difficult to obtain the correct read numbers for each allele using the publicly available RNA-Seq analysis software because read mapping is not performed easily under 100% matching conditions, and in such cases of reduced stringency, there is a possibility that reads from other highly similar HLA genes or alleles may be mapped unexpectedly. Therefore, we focused only on coding regions (exons 2 and 3 of class I loci and exons 2 of class II loci) that can be divided easily into each polymorphic phase and the reads mapped to sequence references at 100% matching condition. Even when using our stringent mapping conditions, 72 different HLA genotypes were excluded from the subsequent analysis due to the partial mapping of the reads. In addition, to better standardize our methods, we excluded homozygous genotypes from our analysis of expressed RNA levels, although the HLA expression levels in homozygotes are approximately double those observed in heterozygotes (20). Therefore, correcting the sequence read numbers using stringent mapping conditions in our study increased the accuracy for quantitating the RNA levels expressed for each allele.

The quantitative results of RNA sequence profiles expressed by 12 HLA loci of PBMC and UCB (**Figure 1** and **Figure S4**) in our study are consistent with the findings of others (58, 59) that the RNA levels expressed by the class I locus usually are higher than by the class II locus. While the RNA expression levels at the class I loci were in the order of *HLA-B* > *HLA-C* > *HLA-A*, the RNA expression levels at the class II locus showed lower locus-specific expression levels with the following relative order of *HLA-DRA* > *HLA-DRB1* > *HLA-DPB1* > *HLA-DPA1* > *HLA-DQB1* > *HLA-DQA1* > *HLA* = *DRB4* > *HLA* = *DRB5* > *HLA-DRB3* (**Figure 1** and **Table 1**). The difference that we found between the different HLA loci, except for *HLA-DRB3/DRB4/DRB5*, correlated well ($R^2 = 0.9135$) with the Illumina bodyMap2 transcriptome analysis (http://www.ensembl.info/2011/05/24/human-bodymap-2-0-data-from-illumina/) using white blood cells (NCBI BioProject Accession: PRJEB2445) (**Figure S5**). In addition, the RNA expression levels of the class I loci in our study were very similar to those of locus-specific real-time PCR using blood leukocytes (59). However, in the previously reported RNA-Seq, the RNA expression level at the class I loci were *HLA-A* ≈ *HLA-B* > *HLA-C*, and the RNA expression level at the class II loci were higher for the alpha chain loci than for the beta chain loci such as *HLA-DRA* > *HLA-DRB1* > *HLA-DQA1* > *HLA-DPA1* > *HLA-DQB1* > *HLA-DPB1* (43). In the RNA-Seq analysis of lymphoblastoid cell lines (LCL) (60), the RNA expression levels at the class I loci showed similar shapes with those of LCL cell lines, JY, and Pala (37). However, we found that the RNA levels expressed by some shared alleles of PBMC and UCB were significantly different (**Figure S4B** and **Table S7**). Since UCB includes stem cells and progenitor cells such as hematopoietic stem cells, mesenchymal stem cells, and vascular endothelial progenitor cells (61), the differences between PBMC and UCB for the RNA levels expressed at the same alleles probably reflect the different cell types. Our results for the RNA levels expressed by *HLA-C*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DPA1* are consistent with the results of Aguiar et al. (43). However, we also observed some inconsistencies; for example, the levels of RNA expressed by the *HLA-A*24* alleles (*A*24:02* and *A*24:20*) in our experiment are different from those previously obtained by quantitative PCR (qPCR) (34). This variability may reflect differences in methodology, cell types (PBMC vs. B lymphocytes), and population ethnicity. It would be important, therefore, to expand our analyses to subpopulations of cells to further understand the basis for allelic RNA-level differences.

We also noted from global comparison of the data that interallelic differences in the RNA levels are less pronounced at the *HLA-A*, *HLA-B*, and *HLA-C* class I loci compared to class II loci such as *HLA-DQA1* and *HLA-DQB1*. Similarly, variations among individuals with identical alleles appear to be smaller at the class I genes compared to the class II genes. Therefore, in contrast to some of the class II genes with high degrees of variations in the RNA levels, expression of the class I genes might have evolved so that their expression levels are maintained to be relatively constant among different alleles and different individuals.

Polymorphisms located in the 5′ promoter region and 3′UTR of HLA genes are known to be associated with variation in HLA expression levels (21, 32–36). Regulatory *cis* and *trans* polymorphisms that affect transcriptional regulation also are involved in susceptibility to complex, multifactorial diseases (13). In this regard, our RNA sequencing method could help to evaluate the role of polymorphisms in the *cis* and *trans* regions of the HLA genes and allele-specific regulation of HLA gene expression because our capture probes were designed so that they covered full-length HLA regions including introns and 5′ and 3′UTR regions. These polymorphic HLA UTR sequences are of interest because some of them are targets for microRNA that regulate the protein and cell surface expression of the HLA genes (17, 19, 21). The capture RNA-Seq method may provide further and more comprehensive data, which could lead to a better understanding of the molecular mechanism or polymorphisms that regulate HLA RNA expression levels. This could be developed in ways to better manage HLA-associated diseases or transplantation outcomes. For example, high expression alleles of *HLA-DPB1* were reported to be a risk effect for acute GVHD (25) (see also above), and therefore, reducing their expression levels prior to transplantation might help to reduce the risk of developing acute GVHD.

Of the 2,175 alleles at 12 HLA loci of 161 PBMC samples, we found on average that the allelic expression differences ranged from 1.3-fold for *HLA-B* to 5.8-fold for *HLA-DQB1* (**Figure 2** and **Table 2**). Especially, large differences were observed in *HLA-DQA1* and *HLA-DQB1* loci (**Figure S3B**).

Restrictions in DQA1/DQB1 heterodimer pairing in which DQA1*01 proteins (low expression) only pair with DQB1*05 or DQB1*06 proteins (high expression) and DQA1*03 proteins (high expression) could pair with DQB1*02, DQB1*03 and DQB1*04 proteins (low expression) were observed in the 17th International Histocompatibility and Immunogenetics Workshop and Workshop Conference (IHIW) NGS HLA genotyping data (http://17ihiw.org/17th-ihiw-ngs-hla-data/) (62). Therefore, it appears that the *DQA1-DQB1* linkage disequilibrium patterns could result from structural interactions of heterodimers associated with low/intermediate expression levels. It has been reported that specific allelic combinations of HLA-*DQA1*, *HLA-DQB1*, and *HLA-DRB1* genes influence autoimmune disease predisposition, and, furthermore, their expression levels may also correlate with causes of the disease.

In several previous studies of HLA gene expression, the HLA genes were genotyped only to the field-1 level (allele lineage level) (21, 34, 43). However, expression studies of HLA genes genotyped only to the level of an allele lineage can miss the diversity that is more evident at the field-3 level. We compared the RNA expression levels among the classified alleles at least up to the field-3 level and classified the RNA expression levels of different alleles from the same HLA locus into distinct high and low groups. For example, we classified the RNA expression levels of 15 *HLA-A* alleles for high and low groups ($P = 1.7 \times 10^{-19}$); *A*26:01:01* and *A*26:02:01* were classified into a high expression group and *A*26:03:01* into a low expression group (**Figure S3E**). In *HLA-DPB1*, a strong significant difference ($P = 7.8 \times 10^{-8}$) was observed in the expression levels of *DPB1*04:01* and *DPB1*04:02*. Allelic differences at the field-3 level were also observed for other HLA loci such as *DQA1*01:03:01* vs. *DQA1*01* group alleles, *DQB1*06:01:01* vs. *DQB1*06* group alleles, and *DRB4*01:03:02* vs. *DRB4*01:03:01* (**Figure S3B** and **Table S5**). Thus, a new nomenclature for HLA alleles that is based on DAE results might be useful, especially in the investigation of transplantation outcome, infections, autoimmunity, cancer, and drug adverse effects. In addition, *DRB3*02:02:01* broadly linked to *HLA-DRB1* alleles and all *HLA-DRB1* types (DR3, DR11, DR12, DR13, and DR14) with an ∼2-fold difference was observed in the expression levels between the *DRB1*14:54:01* linked *DRB3*02:02:01* and *DRB1*14:06:01* linked *DRB3*02:02:01* (**Table S5**). The current NGS methods do not fully cover all the promoter/enhancer and intronic regions, and variations in these segments could determine differences of RNA expression levels that may define the differences between the low and high expressed alleles. Therefore, the newer and most advanced NGS methods should focus on sequencing the segments that determine the RNA expression levels because these possible variations could be important for better understanding the results of associations between HLA alleles, expression, and disease.

Various quantitative methods have been used previously to study specific HLA differential allelic expression including a luciferase reporter assay (19), qPCR using locus-specific primers (21, 38), and flow cytometry using antibodies to measure cellular HLA protein expression (20, 24), but none of these methods permit interallelic comparisons. On the other hand, NGS RNA-Seq (63) and our capture RNA-Seq method enable the RNA expression levels to be compared among the different HLA alleles and enable the detection of null and low expressed HLA alleles (**Figure S3D**). In addition, capture RNA-Seq is a cost-saving method that allowed us to analyze 24 samples in one MiSeq run and to use the same NGS results for genotyping and for quantitating the polymorphic RNA levels. DNA polymorphism analysis of full-length HLA genes (64, 65) and HLA gene expression (58) are usually treated as separate studies. Our singular approach for genotyping RNA as cDNA and also quantitatively measuring allelic expression as sequence read numbers could be a useful new approach for evaluating molecular mechanisms that are involved in abnormal HLA gene expression in cancer cells (31) and in the pathogenesis of autoimmune diseases.

Sequencing-based whole-transcriptome analysis (i.e., RNA-Seq) is a powerful tool to measure gene expression, detect novel transcripts, characterize transcript isoforms, and identify sequence polymorphisms. Here, we described a target enrichment method where a total RNA sample was converted to a sequencing-ready cDNA library and hybridized to a large set of HLA polymorphic RNA-specific biotinylated oligonucleotide capture probes prior to NGS. The resulting sequence data were highly enriched with low expressed alleles that dramatically increased the efficiency of next-generation sequencing and the analysis of allelic expressed RNAs.

## CONCLUSION

This study is the first report of allele expression level differences for 12 classical HLA loci using a novel capture RNA-Seq method. The quantitative DAE data potentially provide information for predicting risks of graft rejections due to abnormally expressed HLA molecules in HCT and for discovering novel pathophysiological mechanisms in HLA-related diseases.

## DATA AVAILABILITY STATEMENT

The novel HLA allele sequence is available in GenBank/DDBJ/ENBL-EBI DNA databases under the Accession Number LC499658. This work described in the article was performed with permission from The Japanese Data Center for Hematopoietic Transplantation (JDCHCT: http://www.jdchct.or.jp/en/outline/). The NGS data will be stored and maintained on a data server at Tokai University School of Medicine for at least 5 years and will be made available to interested parties upon request for validating the findings described in the paper. However, if anybody wants to use the raw NGS data beyond evaluating the current work, it is considered to be secondary usage of the data at JDCHCT (office@jdchct.or.jp), and written permission will need to be obtained from JDCHCT for such usage.

## ETHICS STATEMENT

The study protocol was approved from the institutional review board of the Japan Marrow Donor Program (JMDP) and Tokai University (Application number: 18I-48, 18I-49), and informed consents were obtained from donors in accordance with the Declaration of Helsinki. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

FY, SS, and TS participated in the design of this study. AM, AS, FY, and SI carried out most of the experiments. MM, SK, SM, YK, and YM supported the study. FY, JK, MF-V, MT, SB, SS, and TS analyzed the data and wrote the manuscript. All authors checked the final version of the paper.

## FUNDING

This work was supported by MEXT KAKENHI (No. 16H06502), and by the Practical Research Project for Allergic Diseases and Immunology (Research on Technology of Medical Transplantation) from the Japan Agency for Medical Research and Development (17ek0510022h0001, 18ek0510022h0002 and 19ek0510022h0003). Work in SB's laboratory is supported by grants from the Agence Nationale de la Recherche (ANR) (ANR-11-LABX-0070_TRANSPLANTEX) andMSD-Avenir (AUTOGEN).

## ACKNOWLEDGMENTS

The authors thank the Japan Marrow Donor Program and the Tokai University Cord Blood Bank for providing PBMC and UCB samples.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00941/full#supplementary-material

## REFERENCES

1. Little AM, Parham P. Polymorphism and evolution of HLA class I and II genes and molecules. *Rev Immunogenet.* (1999) 1:105–23.
2. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet.* (2009) 54:15–39. doi: 10.1038/jhg.2008.5
3. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* (2015) 43:D423–31. doi: 10.1093/nar/gku1161
4. Mosaad YM. Clinical role of human leukocyte antigen in health and disease. *Scand J Immunol.* (2015) 82:283–306. doi: 10.1111/sji.12329
5. Zinkernagel RM, Doherty PC. The discovery of MHC restriction. *Immunol Today.* (1997) 18:14–7. doi: 10.1016/S0167-5699(97)80008-4
6. Sasazuki T, Juji T, Morishima Y, Kinukawa N, Kashiwabara H, Inoko H, et al. Effect of matching of class I HLA alleles on clinical outcome after transplantation of hematopoietic stem cells from an unrelated donor. *Japan Marrow Donor Program. N Engl J Med.* (1998) 339:1177–85. doi: 10.1056/NEJM199810223391701
7. Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, et al. The DNA sequence and analysis of human chromosome 6. *Nature.* (2003) 425:805–11. doi: 10.1038/nature02055
8. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens.* (2004) 64:631–49. doi: 10.1111/j.1399-0039.2004.00327.x
9. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Ann Rev Genomics Hum Genet.* (2013) 14:301–23. doi: 10.1146/annurev-genom-091212-153455
10. Fan WL, Shiao MS, Hui RC, Su SC, Wang CW, Chang YC, et al. HLA association with drug-induced adverse reactions. *J Immunol Res.* (2017) 2017:3186328. doi: 10.1155/2017/3186328
11. Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. *J Hum Genet.* (2015) 60:665–73. doi: 10.1038/jhg.2015.102
12. Erlich HA. HLA typing using next generation sequencing: an overview. *Hum Immunol.* (2015) 76:887–90. doi: 10.1016/j.humimm.2015.03.001
13. Knight JC. Regulatory polymorphisms underlying complex disease traits. *J Mol Med.* (2005) 83:97–109. doi: 10.1007/s00109-004-0603-7
14. Ayala FJ, Escalante A, O'Huigin C, Klein J. Molecular genetics of speciation and human origins. *Proc Natl Acad Sci U S A.* (1994) 91:6787–94. doi: 10.1073/pnas.91.15.6787
15. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, et al. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol.* (2003) 20:1377–419. doi: 10.1093/molbev/msg140
16. Cauli A, Dessole G, Fiorillo MT, Vacca A, Mameli A, Bitti P, et al. Increased level of HLA-B27 expression in ankylosing spondylitis patients compared with healthy HLA-B27-positive subjects: a possible further susceptibility factor for the development of disease. *Rheumatology.* (2002) 41:1375–9. doi: 10.1093/rheumatology/41.12.1375
17. Chitnis N, Clark PM, Kamoun M, Stolle C, Brad Johnson F, Monos DS. An expanded role for HLA genes: HLA-B encodes a microRNA that regulates IgA and other immune response transcripts. *Front Immunol.* (2017) 8:583. doi: 10.3389/fimmu.2017.00583
18. Thomas R, Apps R, Qi Y, Gao X, Male V, O'hUigin C, et al. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet.* (2009) 41:1290–4. doi: 10.1038/ng.486
19. Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, et al. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature.* (2011) 472:495–8. doi: 10.1038/nature09914
20. Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, et al. Influence of HLA-C expression level on HIV control. *Science.* (2013) 340:87–91. doi: 10.1126/science.1232685
21. Kulkarni S, Qi Y, O'HUigin C, Pereyra F, Ramsuran V, McLaren P, et al. Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. *Proc Natl Acad Sci U S A.* (2013) 110:20705–10. doi: 10.1073/pnas.1312237110
22. Petersdorf EW, Gooley TA, Malkki M, Bacigalupo AP, Cesbron A, Du Toit E, et al. HLA-C expression levels define permissible mismatches in hematopoietic cell transplantation. *Blood.* (2014) 124:3996–4003. doi: 10.1182/blood-2014-09-599969
23. Hofer TP, Frankenberger M, Heimbeck I, Burggraf D, Wjst M, Wright AK, et al. Decreased expression of HLA-DQ and HLA-DR on cells of the monocytic lineage in cystic fibrosis. *J Mol Med.* (2014) 92:1293–304. doi: 10.1007/s00109-014-1200-z
24. Thomas R, Thio CL, Apps R, Qi Y, Gao X, Marti D, et al. A novel variant marking HLA-DP expression levels predicts recovery from

hepatitis B virus infection. *J Virol.* (2012) 86:6979–85. doi: 10.1128/JVI.00406-12

25. Petersdorf EW, Malkki M, O'hUigin C, Carrington M, Gooley T, Haagenson MD, et al. High HLA-DP expression and graft-versus-host disease. *N Engl J Med.* (2015) 373:599–609. doi: 10.1056/NEJMoa1500140
26. Odani T, Yasuda S, Ota Y, Fujieda Y, Kon Y, Horita T, et al. Up-regulated expression of HLA-DRB5 transcripts and high frequency of the HLA-DRB5*01:05 allele in scleroderma patients with interstitial lung disease. *Rheumatology.* (2012) 51:1765–74. doi: 10.1093/rheumatology/kes149
27. Zhang Y, Liu Y, Lu N, Shan NN, Zheng GX, Zhao SM, et al. Expression of the genes encoding human leucocyte antigens-A, -B, -DP, -DQ and -G in gastric cancer patients. *J Int Med Res.* (2010) 38:949–56. doi: 10.1177/147323001003800321
28. Boegel S, Lower M, Bukur T, Sahin U, Castle JC. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *Oncoimmunology.* (2014) 3:e954893. doi: 10.4161/21624011.2014.954893
29. Schuster H, Peper JK, Bosmuller HC, Rohle K, Backert L, Bilich T, et al. The immunopeptidomic landscape of ovarian carcinomas. *Proc Natl Acad Sci U S A.* (2017) 114:E9942–E51. doi: 10.1073/pnas.1707658114
30. Paulson KG, Tegeder A, Willmes C, Iyer JG, Afanasiev OK, Schrama D, et al. Downregulation of MHC-I expression is prevalent but reversible in Merkel cell carcinoma. *Cancer Immunol Res.* (2014) 2:1071–9. doi: 10.1158/2326-6066.CIR-14-0005
31. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell.* (2017) 171:1259–71 e11. doi: 10.1016/j.cell.2017.10.001
32. Ting JP, Trowsdale J. Genetic control of MHC class II expression. *Cell.* (2002) 109(Suppl):S21–33. doi: 10.1016/S0092-8674(02)00696-7
33. Reith W, LeibundGut-Landmann S, Waldburger JM. Regulation of MHC class II gene expression by the class II transactivator. *Nat Rev Immunol.* (2005) 5:793–806. doi: 10.1038/nri1708
34. Ramsuran V, Kulkarni S, O'hUigin C, Yuki Y, Augusto DG, Gao X, et al. Epigenetic regulation of differential HLA-A allelic expression levels. *Hum Mol Genet.* (2015) 24:4268–75. doi: 10.1093/hmg/ddv158
35. Ramsuran V, Hernandez-Sanchez PG, O'hUigin C, Sharma G, Spence N, Augusto DG, et al. Sequence and phylogenetic analysis of the untranslated promoter regions for HLA class I genes. *J Immunol.* (2017) 198:2320–9. doi: 10.4049/jimmunol.1601679
36. Kaur G, Gras S, Mobbs JI, Vivian JP, Cortes A, Barber T, et al. Structural and regulatory diversity shape HLA-C protein expression levels. *Nat Commun.* (2017) 8:15924. doi: 10.1038/ncomms15924
37. Johnson DR. Differential expression of human major histocompatibility class I loci: HLA-A, -B, and -C. *Hum Immunol.* (2000) 61:389–96. doi: 10.1016/S0198-8859(99)00186-X
38. Bettens F, Brunet L, Tiercy JM. High-allelic variability in HLA-C mRNA expression: association with HLA-extended haplotypes. *Genes Immun.* (2014) 15:176–81. doi: 10.1038/gene.2014.1
39. van Essen TH, van Pelt SI, Bronkhorst IH, Versluis M, Nemati F, Laurent C, et al. Upregulation of HLA expression in primary uveal melanoma by infiltrating leukocytes. *PLoS One.* (2016) 11:e0164292. doi: 10.1371/journal.pone.0164292
40. Small HY, Akehurst C, Sharafetdinova L, McBride MW, McClure JD, Robinson SW, et al. HLA gene expression is altered in whole blood and placenta from women who later developed preeclampsia. *Physiol Genomics.* (2017) 49:193–200. doi: 10.1152/physiolgenomics.00106.2016
41. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep.* (2018) 8:4781. doi: 10.1038/s41598-018-23226-4
42. Deelen P, Zhernakova DV, de Haan M, van der Sijde M, Bonder MJ, Karjalainen J, et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* (2015) 7:30. doi: 10.1186/s13073-015-0152-4
43. Aguiar VRC, Cesar J, Delaneau O, Dermitzakis ET, Meyer D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLos Genet.* (2019) 15:e1008091. doi: 10.1371/journal.pgen.1008091
44. Segawa H, Kukita Y, Kato K. HLA genotyping by next-generation sequencing of complementary DNA. *BMC Genomics.* (2017) 18:914. doi: 10.1186/s12864-017-4300-7
45. Johansson T, Yohannes D, Koskela S, Partanen J, Saavalainen P. HLA RNAseq reveals high allele-specific variability in mRNA expression. *bioRxiv.* (2018). doi: 10.1101/413534
46. Mohlke KL, Erdos MR, Scott LJ, Fingerlin TE, Jackson AU, Silander K, et al. High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc Natl Acad Sci U S A.* (2002) 99:16928–33. doi: 10.1073/pnas.262661399
47. Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, et al. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat Protoc.* (2014) 9:989–1009. doi: 10.1038/nprot.2014.058
48. Morishima Y, Kashiwase K, Matsuo K, Azuma F, Morishima S, Onizuka M, et al. Biological significance of HLA locus matching in unrelated donor bone marrow transplantation. *Blood.* (2015) 125:1189–97. doi: 10.1182/blood-2014-10-604785
49. Suzuki S, Ranade S, Osaki K, Ito S, Shigenari A, Ohnuki Y, et al. Reference grade characterization of polymorphisms in full-length HLA class I and II genes with short-read sequencing on the ION PGM system and long-reads generated by single molecule, real-time sequencing on the PacBio platform. *Front Immunol.* (2018) 9:2294. doi: 10.3389/fimmu.2018.02294
50. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res.* (2002) 12:656–64. doi: 10.1101/gr.229202
51. Shiina T, Suzuki S, Kulski JK. MHC Genotyping in human and nonhuman species by PCR-based next-generation sequencing. In: Kulski JK, editor. *Next Generation Sequencing.* Croatia: Intech (2016). p. 81–109.
52. Holcomb CL, Rastrou M, Williams TC, Goodridge D, Lazaro AM, Tilanus M, et al. Next-generation sequencing can reveal in vitro-generated PCR crossover products: some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. *Tissue Antigens.* (2014) 83:32–40. doi: 10.1111/tan.12269
53. Ozaki Y, Suzuki S, Kashiwase K, Shigenari A, Okudaira Y, Ito S, et al. Cost-efficient multiplex PCR for routine genotyping of up to nine classical HLA loci in a single analytical run of multiple samples by next generation sequencing. *BMC Genomics.* (2015) 16:318. doi: 10.1186/s12864-015-1514-4
54. Ozaki Y, Suzuki S, Shigenari A, Okudaira Y, Kikkawa E, Oka A, et al. HLA-DRB1, -DRB3, -DRB4 and -DRB5 genotyping at a super-high resolution level by long range PCR and high-throughput sequencing. *Tissue Antigens.* (2014) 83:10–6. doi: 10.1111/tan.12258
55. Britten AC, Mijovic CH, Barnett AH, Kelly MA. Differential expression of HLA-DQ alleles in peripheral blood mononuclear cells: alleles associated with susceptibility to and protection from autoimmune type 1 diabetes. *Int J Immunogenet.* (2009) 36:47–57. doi: 10.1111/j.1744-313X.2008.00823.x
56. Yamazaki T, Umemura T, Joshita S, Yoshizawa K, Tanaka E, Ota M. A cis-eQTL of HLA-DPB1 affects susceptibility to type 1 autoimmune hepatitis. *Sci Rep.* (2018) 8:11924. doi: 10.1038/s41598-018-30406-9
57. Fleischhauer K. Immunogenetics of HLA-DP–a new view of permissible mismatches. *N Engl J Med.* (2015) 373:669–72. doi: 10.1056/NEJMe1505539
58. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics.* (2014) 13:397–406. doi: 10.1074/mcp.M113.035600
59. Garcia-Ruano AB, Mendez R, Romero JM, Cabrera T, Ruiz-Cabello F, Garrido F. Analysis of HLA-ABC locus-specific transcription in normal tissues. *Immunogenetics.* (2010) 62:711–9. doi: 10.1007/s00251-010-0470-z
60. Lappalainen T, Sammeth M, Friedlander MR, Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* (2013) 501:506–11. doi: 10.1038/nature12531
61. Forraz N, McGuckin CP. The umbilical cord: a rich and ethical stem cell source to advance regenerative medicine. *Cell Proliferation.* (2011) 44(Suppl 1):60–9. doi: 10.1111/j.1365-2184.2010.00729.x

62. Kwok WW, Kovats S, Thurtle P, Nepom GT. HLA-DQ allelic polymorphisms constrain patterns of class II heterodimer formation. *J Immunol.* (1993) 150:2263–72.

63. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* (2009) 10:57–63. doi: 10.1038/nrg2484

64. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, et al. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. *Tissue Antigens.* (2012) 80:305–16. doi: 10.1111/j.1399-0039.2012.01941.x

65. Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proc Natl Acad Sci U S A.* (2012) 109:8676–81. doi: 10.1073/pnas.1206614109

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Immunology

PERSPECTIVE
published: 21 August 2020
doi: 10.3389/fimmu.2020.01881

# Single Nucleotide Polymorphism in _KIR2DL1_ Is Associated With HLA-C Expression in Global Populations

**_Luciana de Brito Vargas[1], Renata M. Dourado[1], Leonardo M. Amorim[1], Brenda Ho[2], Verónica Calonga-Solís[1], Hellen C. Issler[1], Wesley M. Marin[2], Marcia H. Beltrame[1], Maria Luiza Petzl-Erler[1], Jill A. Hollenbach[2] and Danillo G. Augusto[1,2]*_**

_[1] Programa de Pós-Graduação em Genética, Departamento de Genética, Universidade Federal do Paraná, Curitiba, Brazil, [2] Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, United States_

Regulation of NK cell activity is mediated through killer-cell immunoglobulin-like receptors (KIR) ability to recognize human leukocyte antigen (HLA) class I molecules as ligands. Interaction of KIR and HLA is implicated in viral infections, autoimmunity, and reproduction and there is growing evidence of the coevolution of these two independently segregating gene families. By leveraging _KIR_ and _HLA-C_ data from 1000 Genomes consortium we observed that the _KIR2DL1_ variant _rs2304224*T_ is associated with lower expression of HLA-C in individuals carrying the ligand HLA-C2 ($p = 0.0059$). Using flow cytometry, we demonstrated that this variant is also associated with higher expression of KIR2DL1 on the NK cell surface ($p = 0.0002$). Next, we applied next generation sequencing to analyze _KIR2DL1_ sequence variation in 109 Euro and 75 Japanese descendants. Analyzing the extended haplotype homozygosity, we show signals of positive selection for _rs4806553*G_ and _rs687000*G,_ which are in linkage disequilibrium with _rs2304224*T_. Our results suggest that lower expression of HLA-C2 ligands might be compensated for higher expression of the receptor KIR2DL1 and bring new insights into the coevolution of _KIR_ and _HLA_.

**Keywords: NK cells, KIR, natural selection, linkage disequilibrium, coevolution, expression, population genetics**

OPEN ACCESS

**_Edited by:_**
_Martin Maiers,_
_National Marrow Donor Program,_
_United States_

**_Reviewed by:_**
_Julian Vivian,_
_Monash University, Australia_
_Stephen K. Anderson,_
_National Cancer Institute at Frederick,_
_United States_

**_*Correspondence:_**
_Danillo G. Augusto_
_danillo@augusto.bio.br_

**_Specialty section:_**
_This article was submitted to_
_NK and Innate Lymphoid Cell Biology,_
_a section of the journal_
_Frontiers in Immunology_

**_Received:_** _22 May 2020_
**_Accepted:_** _13 July 2020_
**_Published:_** _21 August 2020_

**_Citation:_**
_Vargas LB, Dourado RM, Amorim LM, Ho B, Calonga-Solís V, Issler HC, Marin WM, Beltrame MH, Petzl-Erler ML, Hollenbach JA and Augusto DG (2020) Single Nucleotide Polymorphism in KIR2DL1 Is Associated With HLA-C Expression in Global Populations. Front. Immunol. 11:1881. doi: 10.3389/fimmu.2020.01881_

## INTRODUCTION

The _killer cell immunoglobulin-like receptor_ (KIR) genes on chromosome 19 encode receptors that interact with a subset of human leukocyte antigen (HLA) class I molecules, encoded by genes on chromosome 6, to regulate NK cell cytotoxicity against infected and neoplastic cells (1–3). In fact, combinations of variants of _KIR_ and _HLA_ have been repeatedly associated with autoimmune disease (4–6), cancer (7, 8), viral infections (9, 10), and are also implicated in reproduction (11–14). As a result, the interaction of KIR and HLA is relevant to fitness and survival and candidate for evolutionary studies (15).

KIR recognize subsets of HLA-A (A3, A11, and Bw4), HLA-B (Bw4 and Bw6), and HLA-C (C1 and C2) molecules (16). _HLA-C_ appears to have had a great impact on _KIR_ evolution, driving the expansion of lineage III KIR, which are the receptor lineage that recognize HLA-C (17, 18). The dimorphism in position 80 of HLA-C defines HLA-C1 ($80^{\text{Asn}}$) and HLA-C2 ($80^{\text{Lys}}$)

and confers differential specificity to KIR. Among all ligands, the interaction between KIR2DL1 and HLA-C2 is responsible for the strongest regulatory signal and HLA-C seems to act as the main educator of NK cells (19, 20).

Worldwide studies demonstrate coordinated frequencies of *KIR* and *HLA* in populations. In a comprehensive study consisting of 30 populations, Single et al. (21) found that increasing frequencies of activating KIR are correlated with decreased frequencies of their respective HLA ligands. On the other hand, Hollenbach et al. (22) showed positive correlation between the presence of KIR2DL3 and the presence of HLA-C1 in 105 worldwide populations. A strong and negative correlation of *KIR* gene-content haplotype *A* and HLA-C2, a pair which is associated with increased risk of pre-eclampsia, was found in eight populations from European, African, and Asian ancestries (11). Moreover, there is extensive evidence of balancing selection maintaining diversity in *KIR* genes (23–25). *KIR* and *HLA* segregate independently and there are no reports of gametic association between these two gene families. Here, we show that a single nucleotide polymorphism (SNP) in *KIR2DL1* is associated with expression levels of the KIR2DL1 receptor on the cell surface and also with HLA-C expression.

## RESULTS

### *KIR2DL1* Variant *rs2304224*T* Is Associated With Lower Expression Levels of HLA-C

To search for possible signals of coevolution between *KIR* and *HLA*, we evaluated if variants in inhibitory KIR that bind to HLA-C could be associated with HLA-C expression levels in global populations. We leveraged the public sequencing information available for all populations in the 1000 Genomes Project (1KGP) (26) and retrieved the genotypic data available for SNPs located within *KIR2DL1* and *KIR2DL23* (*rs2304224, rs11673144, rs12982263, rs34721508, rs35719984,* and *rs35861855*) in 955 individuals of various ancestries. We also obtained *HLA* genotyping data available for those individuals (27).

Subsequently, we used previously published data of HLA-C expression levels (28) and imputed the expression for each *HLA-C* genotype in the 1KGP cohort. The variant *rs2304224*T* was associated with lower HLA-C expression levels in individuals *HLA-C1/C2* ($p = 0.0420$) and *HLA-C2/C2* ($p = 0.0059$), but not in *HLA-C1/C1* individuals ($p = 0.0740$; **Figure 1A** and **Supplementary Figure 1**). This variant is in position 13 of exon 1 and causes a phenylalanine to valine change in the KIR2DL1 signal peptide. We replicated these results by imputing the HLA expression in an independent panel of 308 Brazilians Euro-descendants for which *HLA* genotyping data was available, and we sequenced the first exon of *KIR2DL1* to genotype *rs2304224* ($p = 0.0107$; **Figure 1B**).

To demonstrate that our approach to impute the HLA-C expression is predictive of the cell surface expression *in vivo*, we measured the HLA-C surface levels of fresh CD3$^+$ cells in 30 individuals using flow cytometry and compared to the imputed values. We found a correlation of $r = 0.62$, $p < 0.0001$ (**Supplementary Figure 2**).

### *rs2304224*T* Is also Associated With Higher Surface Expression Levels of *KIR2DL1*

We sought to investigate if the variant *rs2304224*T* in *KIR2DL1* was associated with KIR2DL1 surface expression. We used flow cytometry to quantify both the abundance of KIR2DL1 on the surface of NK cells (median fluorescence intensity, MFI) as well as the percentage of NK cells expressing KIR2DL1 on their surface (KIR2DL1$^+$ NK cell), and also interrogated if copy number variation of *KIR2DL1* affects surface expression. Although borderline, we did not find significant differences of expression levels in individuals carrying one copy (hemizygous) or two copies (homo- or heterozygous) of *KIR2DL1*$^+$ ($p = 0.0594$; **Supplementary Figure 3A**). However, the number of KIR2DL1+ NK cells was 2.16-fold higher in individuals carrying two copies ($p = 0.0001$; **Supplementary Figure 3B**). For all KIR2DL1 expression analyses, we used copy number of *KIR2DL1* as covariant in the regression model.

We observed that the allele *rs2304224*T,* associated with decreased HLA-C expression, was also associated with 1.54-fold increase of the KIR2DL1 surface expression ($p = 0.0002$) and a 1.41-fold increase of KIR2DL1$^+$ NK cells ($p = 0.03$; **Figures 1C,D**). The median expression of each KIR allotype is shown in **Supplementary Figure 4**. We also observed that KIR2DL1 expression was decreased in individuals homozygous for the presence of the C2 ligand (*C2/C2*, $p = 0.007$; **Figure 1E**).

### Signals of Positive Selection for *KIR2DL1* Variants in Linkage Disequilibrium With *rs2304224*

We next analyzed the entire *KIR2DL1* gene in a subset of 109 Euro-descendants and 75 Japanese descendants sequenced using our custom next generation sequencing method (29). In Euro descendants, we observed low correlation but strong linkage disequilibrium (LD) between *rs2304224* and three other variants (**Supplementary Figures 5A,B**). The first variant is at position −406 upstream of the *KIR2DL1* gene (*rs4806553*, $D^{'} = 0.99$, $r^2 = 0.18$, $p < 10^{-8}$). The other variants are located within the coding region, in exon 4 (*rs687000*, $D^{'} = 0.99$, $r^2 = 0.52$, $p < 10^{-12}$) and exon 7 (*rs34721508*, $D^{'} = 0.99$, $r^2 = 0.24$, $p < 10^{-3}$). Weaker LD was observed for the same variants in Japanese descendants (**Supplementary Figures 5C,D**). Frequencies for all SNPs in both populations are given in **Supplementary Table 1**. Moreover, the frequency of HLA-C2 in our Japanese-descendant cohort was 10.3% while in Euro-descendants it was 40.9%.

We searched for signals of population specific selection, for both Euro and Japanese descendants, by estimating the extended haplotype homozygosity (EHH) using *rs2304224* and also variants in significant LD with it as focal SNPs. The bifurcation patterns are consistent with positive selection increasing frequencies of the haplotype more rapidly than they could be broken by genetic recombination. Signals of positive selection were observed for the derived allele *rs4806553*G* in

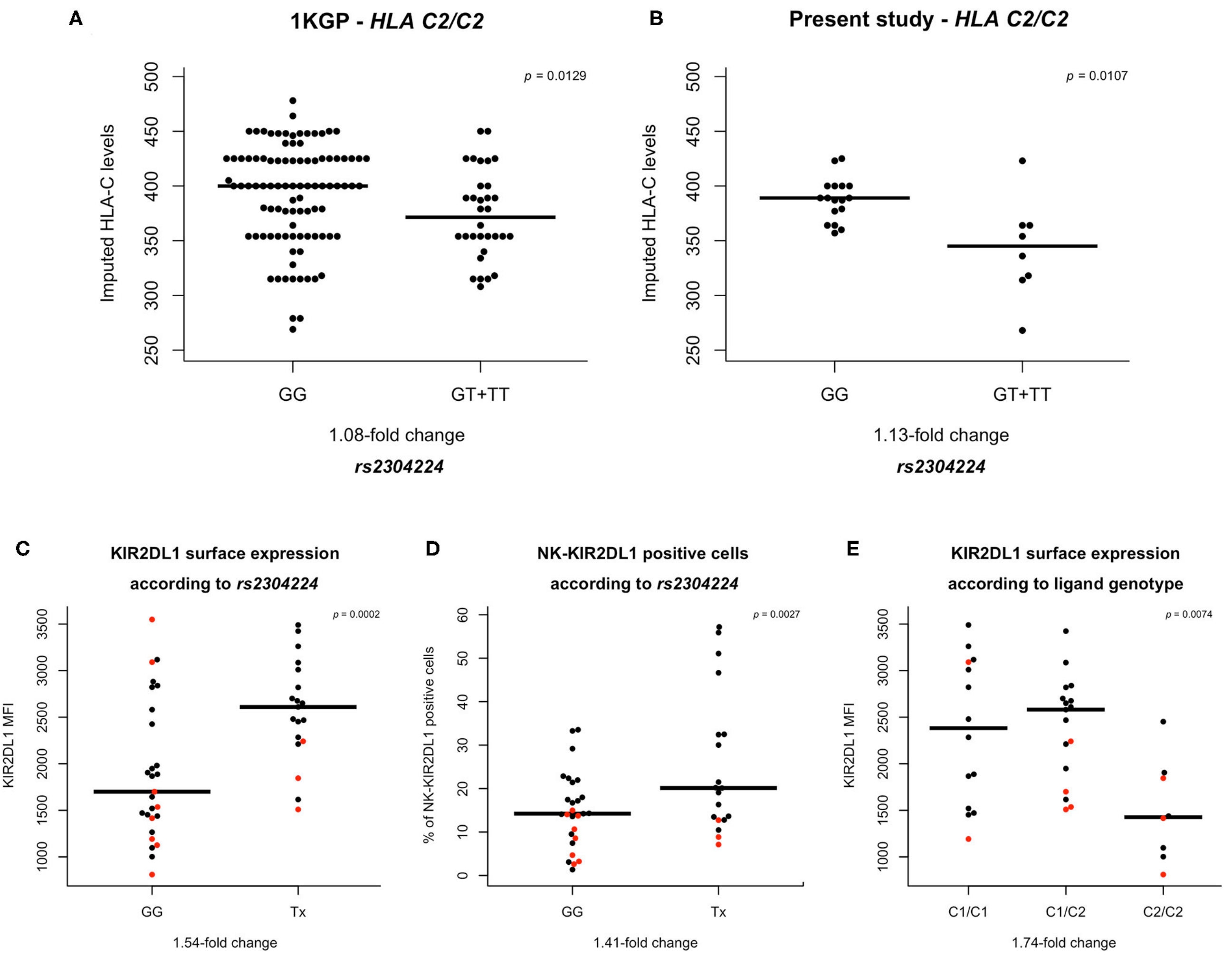


**FIGURE 1** | HLA-C and KIR2DL1 expression are associated with genetic variants. **(A,B)** *rs2304224* in *KIR2DL1* marks *in silico* HLA-C surface expression (28) in two different cohorts. The presence of allele *rs2304224*T* marks lower HLA-C expression in **(A)** 130 *C2/C2* homozygotes out of 955 individuals from 1000 genomes consortium and **(B)** 25 *C2/C2* homozygotes out of 308 Euro-Brazilians from Curitiba (present study). **(C)** Higher KIR2DL1 surface expression and **(D)** increased presence on NK cells are also associated with the variant *rs2304224*T* ($p = 0.0002$ and $p = 0.0027$, respectively). **(E)** *HLA-C* genotype is associated to KIR2DL1 surface expression ($p = 0.0074$). There is no difference in expression, however, between homozygotes *C1/C1* and heterozygotes *C1/C2* ($p = 0.44$). Homozygosity for *C2/C2*, on the other hand, is associated with lower KIR2DL1 surface expression than in *C1/C1* ($p = 0.0031$) and *C1/C2* ($p = 0.0016$). Each dot in the graphs represents one individual. Red dots indicate hemizygosity for *KIR2DL1*. Median values are shown in horizontal lines and statistical significance is indicated in the top right corners of each plot.

Japanese but not in Euro-descendants (**Figures 2A,C**). Strong signals of positive selection were also observed for the derived allele *rs687000*G* in both Euro and Japanese descendants (**Figures 2B,D**).

## DISCUSSION

Previous results show that *cis* polymorphisms associated with *HLA-C* expression do not associate with NK cell activity (30), despite the compelling evidence that KIR-HLA are coevolving as an integrated system (11, 16, 21, 22). Here, we show evidence of coevolution of *KIR* and *HLA* by identifying a variant in *KIR2DL1* that was associated with surface expression of the ligand HLA-C2 in worldwide populations. The allele *rs2304224*T* was associated with lower expression of imputed HLA-C surface expression in 995 individuals from 1KGP and also in an independent cohort of 308 Brazilian Euro-descendants. The association was only observed in individuals carrying at least one copy of HLA-C2, which suggests an orchestrated and refined evolution between these two systems. Although the antibody used in this study (DT9) cross reacts with HLA-E, it has been demonstrated that its binding represents the surface expression of HLA-C (28, 31) and also is correlated with mRNA expression levels of *HLA-C* measured by quantitative PCR (32). Therefore, our direct measurement of HLA-C expression in 30 individuals demonstrates that imputing HLA expression based on previously

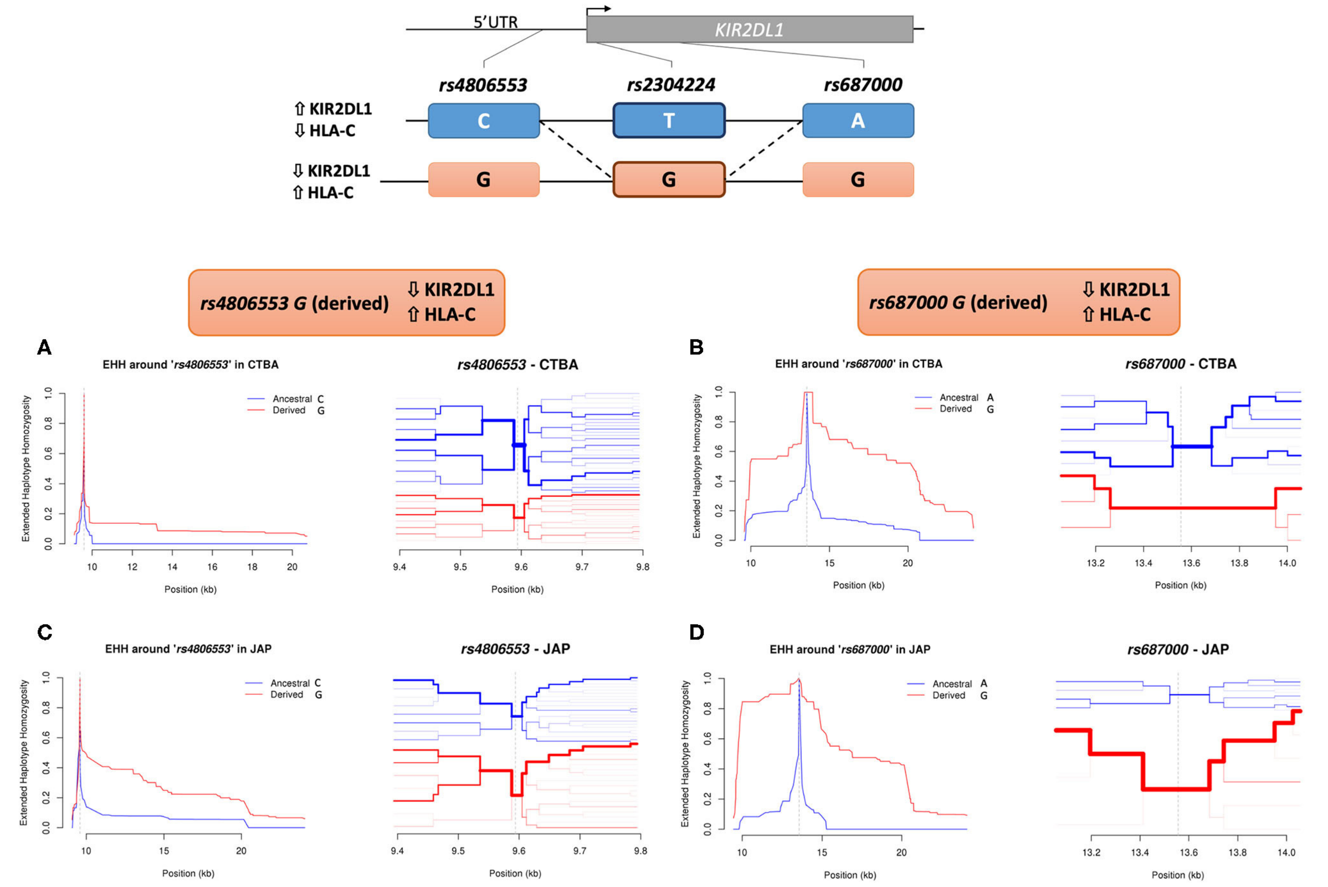


**FIGURE 2 |** Extended haplotype homozygosity (EHH) in *KIR2DL1*. The extended homozygosity analysis is based on the premise that advantageous alleles increase in frequency at a higher pace than the local recombination rate breaks down the haplotypes in which these alleles are located. Therefore, alleles marking regions with elevated extended homozygosity are possibly under recent positive selection. Here we identify extended haplotypes surrounding *KIR2DL1* variants *rs4806553* and *rs687000.* The possible haplotypes of *rs4806553* and *rs687000* in relation to *rs2304224* are represented at the top of the image. The continuous line represents the most common configuration between two variants, and the dashed line represent less frequent configurations. On the left of each haplotype, arrows indicate higher or lower expression of KIR2DL1 and HLA-C, as associated with *rs2304224* alleles *G* or *T*. A representation of the genomic organization of *KIR2DL1* with the indicated location of the three variants is represented above. **(A)** EHH graph of decay in homozygosity (left) and furcation plot (right) for *rs4806553* in Euro-Brazilians. The graph shows little to no difference between ancestral *rs4806553*C* (blue) and derived *rs4806553*G* (red) alleles in Euro-Brazilians. **(C)** EHH graph of decay in homozygosity (left) and furcation plot (right) for *rs4806553* in Japanese. In Japanese, elevated homozygosity is associated with derived allele *rs4806553*G* (red). **(B)** EHH graph of decay in homozygosity (left) and furcation plot (right) for *rs687000* in Brazilians with European ancestry and **(D)** Brazilians with Japanese ancestry. Elevated homozygosity associated with derived allele *rs687000*G* (red) is consistent with the selective sweep model, in which recent positive selection sweeps the diversity on nearby loci. Vertical dotted lines indicate the position of the core SNP. The thickness of each branch in the furcation plot is determined by haplotype frequency.

published data is predictive of the expression observed on the surface of fresh blood cells.

It is also interesting that the same allele *rs2304224*T* is associated with higher expression of the receptor KIR2DL1 in NK cells and also present in the high expressing *KIR2DL1*002*. The SNP *rs2304224* in exon 1 causes a non-synonymous substitution of valine (allele *G*) to phenylalanine (allele *T*) in the signal peptide. The hydrophobicity of the signal peptide can influence protein retention in the cytosol (33). According to the Wimley-White interfacial hydrophobicity scale (34), valine has a free energy of transfer of 0.07 ΔG from water to bilayer, and the free energy of phenylalanine is −1.13 ΔG. The lower and negative value of phenylalanine indicates this transference is more favorable, and therefore, *rs2304224*T* may increase protein availability in the membrane. This could explain the increased KIR2DL1 expression associated with *rs2304224*T*.

The patterns that we observed for the expression of KIR2DL1 allotypes (**Supplementary Figure 4**) are consistent with previous studies (20, 35–37). Our results showing that copy number of *KIR2DL1* affects the quantity of KIR2DL1$^+$ NK cells corroborate those by Béziat et al. (37). On the other hand, the lack of significant association that we observed between *KIR2DL1* copy number and the abundance of expression on the cell surface reinforces the idea that copy number does not affect levels of KIR2DL1 as strongly as it affects the proportion of cells expressing the receptor (37). The presence of HLA-C2 was associated with lower expression of surface KIR2DL1, according to our results and of others (35, 38, 39). However, differently from the observations from Le Luduec et al. (38), who observed that the expression of KIR2DL1 is associated to the presence of C2 in a dose dependent manner, we found association only in individuals carrying two copies of *C2*.

We found three SNPs in LD with *rs2304224* ($D' = 0.99$, $0.18 \leq r^2 \leq 0.51$). The low correlation coefficient is explained by difference in the allele frequencies among them. The frequency of the variant *rs2304224**T* is 0.26 in Euro-Brazilians, while the frequency of *rs4806553**C* is 0.67; *rs687000**A* is 0.57; and *rs34721508**C* is 0.86. From the three variants in LD with *rs2304224,* only *rs34721508*, in exon 7, has been previously associated with differential expression levels of KIR2DL1 in transfected cell lines (36). That study showed that cells expressing allotypes with $245^{\text{Cys}}$ have reduced protein stability and are more susceptible to ligand mediated expression down-regulation in comparison to those with $245^{\text{Arg}}$. Interestingly, this variant was also present in the 1KGP dataset, and we did not observe association of *rs34721508* genotypes with HLA-C imputed expression levels ($p = 0.28$). We also demonstrated that there is an additive effect of *rs2304224**T* and *rs34721508**C* on KIR2DL1 expression, which indicates that each has independent effect on the expression of KIR2DL1 (**Supplementary Figure 6**), despite the fact that both these variants are present in the high expressing *KIR2DL1**002* (**Supplementary Table 1**). This observation argues in favor of our approach to expand the analysis of individual SNPs rather than solely analyzing the common combinations of SNPs present in the most frequent *KIR2DL1* alleles.

We applied extended haplotype homozygosity (EHH) analysis to all SNPs in LD with *rs2304224,* using the next generation sequencing data that we generated for a subset of Euro and Japanese descendants. Homozygosity surrounding the derived allele *rs4806553**G* was prominent in the Japanese population, suggesting this allele has been under recent positive selection. Japanese populations are especially interesting because they exhibit the lowest frequency of the HLA-C2 allotype (only 8%) (40) and, accordingly, we report low frequency of C2 also in the Brazilians of Japanese ancestry (10.3%). The low frequency of HLA-C2 could be driving the evolution of *KIR2DL1* in the Japanese population.

The SNP *rs4806553* is located 406 kbp upstream of the *KIR2DL1* gene, in the sequence corresponding to its intermediate promoter (Pro-I), suggested to control protein expression in mature NK cells (41). Moreover, it has been shown that the Pro-I sequence containing allele *rs4806553**C* binds to the transcription factor activator protein-1 (AP1), while *rs4806553**G* abrogates this binding (42). This could potentially explain the higher expression of *KIR2DL1**002*, which contains allele *rs4806553**C*, in comparison to other *KIR2DL1* alleles carrying the variant *rs4806553**G*, such as *KIR2DL1**004* and *KIR2DL1**006* (**Supplementary Figure 4** and **Supplementary Table 1**). Our data suggests that the attenuation of NK inhibition mediated by KIR2DL1 represents an evolutionary advantage and is being favored by positive selection in the Japanese population.

Strong signals of positive selection were observed toward the derived allele *rs687000**G* in both our cohorts. This variant is a synonymous change in exon 4, without apparent impact on regulation of *KIR2DL1* expression. One hypothesis is that *rs687000**G* rose in frequency due to hitchhiking with a nearby variation that was positively selected and eventually fixed. We did not observe signals of positive selection for *rs2304224* and *rs34721508*, which strongly associate with KIR2DL1 expression levels. One possibility is that selection could be favoring specific *KIR2DL1* alleles that carry these variants. In fact, the combination of *rs2304224**G* (neutral)*, rs687000**G* (positively selected), and *rs34721508**C* (neutral) defines *KIR2DL1**003*, the most frequent allele across all populations worldwide (43).

Coevolution of *KIR* and *HLA* is mostly driven by *HLA-C* (20, 44), which encodes a strong educator for $\text{KIR}^+$ NK cells (45, 46). A fine tuning mechanism of NK cell regulation through the cell-specific promoter NK-Pro (47) was recently proposed, in which expression levels of HLA-C during NK cell education combines with expression levels and interaction strength of KIR and HLA in mature NK cells to modulate their selectivity and cytotoxicity (48). KIR2DL1 is the receptor with the highest affinity and avidity to HLA-C, and mediates the strongest NK response (19, 20, 49). Therefore, it is plausible that variation in *KIR2DL1* could be under selection and also that *KIR2DL1* and *HLA-C* are coevolving. Here, we show a *KIR2DL1* variant that is associated with lower expression of KIR2DL1 and inversely associated with higher HLA-C expression in HLA-C2/C2 individuals. This could be an indication that higher levels of the ligand are being compensated by lower expression of the receptor. We also observed evidence of positive selection on KIR2DL1. Our data show that much remains to be understood regarding the mechanisms of the KIR-HLA recognition and evolution. They also bring insights into the evolution of these two systems and suggest that more questions will emerge as we explore more deeply *KIR-HLA* diversity at high resolution.

# MATERIALS AND METHODS

## Samples

We analyzed a cohort of 308 individuals of predominantly European ancestry and 75 individuals of Japanese ancestry from Curitiba, Brazil. About 80% of the population from Curitiba self-reported as Euro-descendant (50), which is in accordance with previous genetic studies (51). For the Japanese descendants, we only included individuals who had two parents or four grandparents born in Japan, with no history of admixture with non-Japanese ancestries. In order to measure KIR2DL1 expression levels, we analyzed fresh blood cells from a subset of 48 Euro-descendants. A subset of 30 of these individuals were included in the HLA-C expression assay. Detailed information about the study design is given in **Supplementary Figure 7**. All individuals were living in Curitiba, Brazil, at the time of blood collection. Median age in the group was 26 years (ranging from 20 to 64) and the male/female ratio was 0.37.

For expression assays, we collected 8 mL of peripheral blood samples and isolated PBMC (peripheral blood mononuclear cells) using Leucosep$^{TM}$ tubes (Greiner Bio-One, Austria), which have a selective membrane for density-based lymphocyte separation, and Ficoll Hypaque (Sigma Aldrich, MO). Isolated PBMC were counted in a Neubauer chamber under an optical microscope. A total of 0.5 × $10^6$ cells were incubated with specific antibodies for KIR2DL1 and HLA-C and analyzed by flow cytometry. Detailed description and gate strategy are shown in **Supplementary Figure 8**.

## *KIR2DL1* and *HLA-C* Genotyping

We initially sequenced exons 1, 4, 5, 7, and 9 to distinguish the main *KIR2DL1* allele groups using the Sanger method (52) in the 48 Euro-descendants included in the expression assay (**Supplementary Figure 9**). The sequences obtained were aligned with reference sequences from IPD-KIR database (43), using the software Mutation Surveyor® (SoftGenetics, PA) and identified manually. Additionally, we sequenced only the exon 1 (containing the variant *rs2304224*) in extra 260 Euro-descendant individuals to increase statistical power for the analysis of *rs2304224*.

We applied quantitative PCR to determine copy number of *KIR2DL1* compared to *KIR3DL3*, which is present in virtually all haplotypes. *KIR2DL1* was amplified in triplicate using one set of primers and the reference gene *KIR3DL3* was amplified using other three sets of primers, each in triplicate, in a total of 12 (4 × 3) reactions per sample. The sequence of all primers used for amplification, sequencing and copy number assay, including those designed in this study as well as those described previously (53–57) are given in **Supplementary Table 2**.

We also sequenced the entire *KIR2DL1* gene in 109 Euro-descendants and 75 Japanese descendants from Curitiba, Brazil. These samples were sequenced using the previously published method for next generation sequencing of *KIR* and *HLA* genes (29) using Illumina platform.

## Data Analysis

Normality of variables was tested using Kolmogorov-Smirnov test, in R package *nortest* (58). Difference in HLA-C expression between *KIR2DL1* SNP genotypes was tested via the Kruskal-Wallis test, using *stats* R (59). *Post-hoc* analysis of Dunn was applied to Kruskal-Wallis results in order to identify pair-wise significant differences between genotypes, in R package *dunn.test* (60). Median HLA-C expression by allele, as defined by Apps et al. (28), was imputed for each allele in an individual, and then summed. The imputation was performed in all 308 Brazilians of European ancestry sequenced for *rs2304224* and 1KGP individuals. Correlation analysis between expected HLA-C expression in CD3$^+$ cells and *in vivo* HLA-C expression in CD3$^+$ cells was calculated with R package *Hmisc* (61). Difference in KIR2DL1 expression according to copy number was tested using Mann-Whitney, in *stats* R (59). Association of KIR2DL1 expression with allotype and *rs2304224* was tested through logistic regression using copy number as a covariate, also in *stats* R. Linkage disequilibrium was estimated using LD function from R package *genetics* (62) and plotted with a modified version of R package *LDheatmap* (63). Median expression graphs were plotted using *base* and *beeswarm* R packages (59, 64).

*KIR2DL1* SNPs obtained from genomic sequence data were phased using fastPHASE, with modified parameters (-T10 -H200). The phased data was used for estimation of extended haplotype homozygosity (EHH) (65) using R package *rehh* (66). Ancestral and derived alleles were defined according to the Database of Single Nucleotide Polymorphisms (dbSNP) (67).

# DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

# ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Brazilian National Human Research Ethics Committee (CONEP), Protocol No. CAAE 02727412.4.0000.0096, in accordance to the Brazilian Federal laws. The patients/participants provided their written informed consent to participate in this study.

# AUTHOR CONTRIBUTIONS

DA designed the study. LV, RD, VC-S, LA, and HI performed Sanger sequencing and genotyping. LV, DA, and BH performed next generation sequencing. LV, RD, and DA performed flow cytometry analysis. LV, DA, and WM analyzed the data. MP-E, JH, and DA contributed with samples and/or reagents. LV and DA drafted the manuscript. All authors contributed to the article and approved the submitted version.

# FUNDING

This work was supported by grants and scholarships from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Araucária and NIH grant U19NS095774.

# ACKNOWLEDGMENTS

We would like to thank all the volunteers who participated in this study and all the members of the Human Molecular Laboratory of the Federal University of Paraná, Brazil, for their unconditional support.

# SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.01881/full#supplementary-material

## REFERENCES

1. Kiessling R, Klein E, Pross H, Wigzell H. "Natural" killer cells in the mouse. II. Cytotoxic cells with specificity for mouse moloney leukemia cells. Characteristics of the killer cell. *Eur J Immunol.* (1975) 5:117–21. doi: 10.1002/eji.1830050209
2. Waggoner SN, Reighard SD, Gyurova IE, Cranert SA, Mahl SE, Karmele EP, et al. Roles of natural killer cells in antiviral immunity. *Curr Opin Virol.* (2016) 16:15–23. doi: 10.1016/j.coviro.2015.10.008
3. Ciccone E, Pende D, Viale O, Di Donate C, Tripodi G, Orengo AM, et al. Evidence of a natural killer (NK) cell repertoire for (allo) antigen recognition: definition of five distinct NK-determined allospecificities in humans. *J Exp Med.* (1992) 175:709–18. doi: 10.1084/jem.175.3.709
4. van der Slik AR, Koeleman BPC, Verduijn W, Bruining GJ, Roep BO, Giphart MJ. KIR in type 1 diabetes: disparate distribution of activating and inhibitory natural killer cell receptors in patients versus HLA-matched control subjects. *Diabetes.* (2003) 52:2639–42. doi: 10.2337/diabetes.52.10.2639
5. Augusto DG, Lobo-Alves SC, Melo MF, Pereira NF, Petzl-Erler ML. Activating KIR and HLA Bw4 ligands are associated to decreased susceptibility to pemphigus foliaceus, an autoimmune blistering skin disease. *PLoS ONE.* (2012) 7:e39991. doi: 10.1371/journal.pone.0039991
6. Nelson GW, Martin MP, Gladman D, Wade J, Trowsdale J, Carrington M. Cutting edge: heterozygote advantage in autoimmune disease: hierarchy of protection/susceptibility conferred by HLA and killer Ig-like receptor combinations in psoriatic arthritis. *J Immunol.* (2004) 173:4273–6. doi: 10.4049/jimmunol.173.7.4273
7. Jobim MR, Jobim M, Salim PH, Portela P, Jobim LF, Leistner-Segal S, et al. Analysis of KIR gene frequencies and HLA class I genotypes in breast cancer and control group. *Hum Immunol.* (2013) 74:1130–3. doi: 10.1016/j.humimm.2013.06.021
8. Middleton D, Diler AS, Meenagh A, Sleator C, Gourraud PA. Killer immunoglobulin-like receptors (KIR2DL2 and/or KIR2DS2) in presence of their ligand (HLA-C1 group) protect against chronic myeloid leukaemia. *Tissue Antigens.* (2009) 73:553–60. doi: 10.1111/j.1399-0039.2009.01235.x
9. Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, et al. Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nat Genet.* (2007) 39:733–40. doi: 10.1038/ng2035
10. Khakoo SI, Thio CL, Martin MP, Brooks CR, Gao X, Astemborski J, et al. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science.* (2004) 305:872–4. doi: 10.1126/science.1097670
11. Hiby SE, Walker JJ, O'Shaughnessy KM, Redman CWG, Carrington M, Trowsdale J, et al. Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. *J Exp Med.* (2004) 200:957–65. doi: 10.1084/jem.20041214
12. Trowsdale J, Moffett A. NK receptor interactions with MHC class I molecules in pregnancy. *Semin Immunol.* (2008) 20:317–20. doi: 10.1016/j.smim.2008.06.002
13. Nakimuli A, Chazara O, Hiby SE, Farrell L, Tukwasibwe S, Jayaraman J, et al. A KIR B centromeric region present in Africans but not Europeans protects pregnant women from pre-eclampsia. *Proc Natl Acad Sci USA.* (2015) 112:845–50. doi: 10.1073/pnas.1413453112
14. Bulmer JN, Lash GE. The role of uterine NK cells in normal reproduction and reproductive disorders. *Adv Exp Med Biol.* (2015) 868:95–126. doi: 10.1007/978-3-319-18881-2_5
15. Parham P. MHC class I molecules and KIRS in human history, health and survival. *Nat Rev Immunol.* (2005) 5:201–14. doi: 10.1038/nri1570
16. Augusto DG, Petzl-Erler ML. KIR and HLA under pressure: evidences of coevolution across worldwide populations. *Hum Genet.* (2015) 134:929–40. doi: 10.1007/s00439-015-1579-9
17. Older Aguilar AM, Guethlein LA, Adams EJ, Abi-Rached L, Moesta AK, Parham P. Coevolution of killer cell Ig-like receptors with HLA-C to become the major variable regulators of human NK cells. *J Immunol.* (2010) 185:4238–51. doi: 10.4049/jimmunol.1001494
18. Parham P, Guethlein LA. Genetics of natural killer cells in human health, disease, and survival. *Annu Rev Immunol.* (2018) 36:519–48. doi: 10.1146/annurev-immunol-042617-053149
19. Stewart CA, Laugier-Anfossi F, Vely F, Saulquin X, Riedmuller J, Tisserant A, et al. Recognition of peptide-MHC class I complexes by activating killer immunoglobulin-like receptors. *Proc Natl Acad Sci USA.* (2005) 102:13224–9. doi: 10.1073/pnas.0503594102
20. Hilton HG, Guethlein LA, Goyos A, Nemat-Gorgani N, Bushnell DA, Norman PJ, et al. Polymorphic HLA-C receptors balance the functional characteristics of KIR haplotypes. *J Immunol.* (2015) 195:3160–70. doi: 10.4049/jimmunol.1501358
21. Single RM, Martin MP, Gao X, Meyer D, Yeager M, Kidd JR, et al. Global diversity and evidence for coevolution of KIR and HLA. *Nat Genet.* (2007) 39:1114–9. doi: 10.1038/ng2077
22. Hollenbach JA, Augusto DG, Alaez C, Bubnova L, Fae I, Fischer G, et al. 16th IHIW: population global distribution of killer immunoglobulin-like receptor (KIR) and ligands. *Int J Immunogenet*. (2013) 40:39–45. doi: 10.1111/iji.12028
23. Augusto DG, Norman PJ, Dandekar R, Hollenbach JA. Fluctuating and geographically specific selection characterize rapid evolution of the human KIR region. *Front Immunol.* (2019) 10:989. doi: 10.3389/fimmu.2019.00989
24. Gendzekhadze K, Norman PJ, Abi-Rached L, Layrisse Z, Parham P. High KIR diversity in Amerindians is maintained using few gene-content haplotypes. *Immunogenetics.* (2006) 58(5–6):474–80. doi: 10.1007/s00251-006-0108-3
25. Nemat-Gorgani N, Edinur HA, Hollenbach JA, Traherne JA, Dunn PPJ, Chambers GK, et al. KIR diversity in Māori and polynesians: populations in which HLA-B is not a significant KIR ligand. *Immunogenetics.* (2014) 66:597–611. doi: 10.1007/s00251-014-0794-1
26. The 1000Genomes Project Consortium. A global reference for human genetic variation. *Nature.* (2015) 526:68–74. doi: 10.1038/nature15393
27. Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. *PLoS ONE.* (2014) 9:e97282. doi: 10.1371/journal.pone.0097282
28. Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, et al. Influence of HLA-C expression level on HIV control. *Science.* (2013) 340:87–91. doi: 10.1126/science.1232685
29. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. *Am J Hum Genet.* (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023
30. Charoudeh HN, Schmied L, Gonzalez A, Terszowski G, Czaja K, Schmitter K, et al. Quantity of HLA-C surface expression and licensing of KIR2DL+ natural killer cells. *Immunogenetics.* (2012) 64:739–45. doi: 10.1007/s00251-012-0633-1
31. Thomas R, Apps R, Qi Y, Gao X, Male V, O'Huigin C, et al. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet.* (2009) 41:1290–4. doi: 10.1038/ng.486
32. Kulkarni S, Qi Y, O'hUigin C, Pereyra F, Ramsuran V, McLaren P, et al. Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. *Proc Natl Acad Sci USA.* (2013) 110:20705–10. doi: 10.1073/pnas.1312237110
33. Zhang L, Leng Q, Mixson AJ. Alteration in the IL-2 signal peptide affects secretion of proteins *in vitro* and *in vivo*. *J Gene Med.* (2005) 7:354–65. doi: 10.1002/jgm.677
34. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Biol.* (1996) 3:842–8. doi: 10.1038/nsb1096-842
35. Dunphy SE, Guinan KJ, Chorcora CN, Jayaraman J, Traherne JA, Trowsdale J, et al. 2DL1, 2DL2 and 2DL3 all contribute to KIR phenotype variability on human NK cells. *Genes Immun.* (2015) 16:301–10. doi: 10.1038/gene.2015.15
36. Bari R, Bell T, Leung WH, Vong QP, Chan WK, Das Gupta N, et al. Significant functional heterogeneity among KIR2DL1 alleles and a pivotal role of arginine245. *Blood.* (2009) 114:5182–90. doi: 10.1182/blood-2009-07-231977
37. Béziat V, Traherne JA, Liu LL, Jayaraman J, Enqvist M, Larsson S, et al. Influence of KIR gene copy number on natural killer cell education. *Blood.* (2013) 121:4703–7. doi: 10.1182/blood-2012-10-461442
38. Le Luduec JB, Boudreau JE, Freiberg JC, Hsu KC. Novel approach to cell surface discrimination between KIR2DL1 subtypes and KIR2DS1 identifies hierarchies in NK repertoire, education, and tolerance. *Front Immunol.* (2019) 10:734. doi: 10.3389/fimmu.2019.00734
39. He Y, Tao S, Ying Y, He J, Zhu F, Lv H. Allelic polymorphism, mRNA and antigen expression of KIR2DL1 in the Chinese Han population. *Hum Immunol.* (2014) 75:245–9. doi: 10.1016/j.humimm.2013.12.005
40. Yawata M, Yawata N, Draghi M, Little A-M, Partheniou F, Parham P. Roles for HLA and KIR polymorphisms in natural killer cell repertoire

selection and modulation of effector function. *J Exp Med.* (2006) 203:633–45. doi: 10.1084/jem.20051884

41. Wright PW, Li H, Huehn A, O'Connor GM, Cooley S, Miller JS, et al. Characterization of a weakly expressed KIR2DL1 variant reveals a novel upstream promoter that controls KIR expression. *Genes Immun.* (2014) 15:440–8. doi: 10.1038/gene.2014.34
42. Li H, Wright PW, McCullen M, Anderson SK. Characterization of KIR intermediate promoters reveals four promoter types associated with distinct expression patterns of KIR subtypes. *Genes Immun.* (2016) 17:66–74. doi: 10.1038/gene.2015.56
43. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* (2015) 43:D423–31. doi: 10.1093/nar/gku1161
44. Nemat-Gorgani N, Hilton HG, Henn BM, Lin M, Gignoux CR, Myrick JW, et al. Different selected mechanisms attenuated the inhibitory interaction of KIR2DL1 with C2 + HLA-C in two indigenous human populations in Southern Africa. *J Immunol.* (2018) 200:2640–55. doi: 10.4049/jimmunol.1701780
45. David G, Djaoud Z, Willem C, Legrand N, Rettman P, Gagne K, et al. Large spectrum of HLA-C recognition by killer Ig–like receptor (KIR)2DL2 and KIR2DL3 and restricted C1 specificity of KIR2DS2: dominant impact of KIR2DL2/KIR2DS2 on KIR2D NK cell repertoire formation. *J Immunol.* (2013) 191:4778–88. doi: 10.4049/jimmunol.1301580
46. Horowitz A, Djaoud Z, Nemat-Gorgani N, Blokhuis J, Hilton HG, Béziat V, et al. Class I HLA haplotypes form two schools that educate NK cells in different ways. *Sci Immunol.* (2016) 1:eaag1672. doi: 10.1126/sciimmunol.aag1672
47. Li H, Ivarsson MA, Walker-Sperling VE, Subleski J, Johnson JK, Wright PW, et al. Identification of an elaborate NK-specific system regulating HLA-C expression. *PLoS Genet.* (2018) 14:e1007163. doi: 10.1371/journal.pgen.1007163
48. Goodson-Gregg FJ, Krepel SA, Anderson SK. Tuning of human NK cells by endogenous HLA-C expression. *Immunogenetics.* (2020) 72:205–15. doi: 10.1007/s00251-020-01161-x
49. Moesta AK, Norman PJ, Yawata M, Yawata N, Gleimer M, Parham P. Synergistic polymorphism at two positions distal to the ligand-binding site makes KIR2DL2 a stronger receptor for HLA-C than KIR2DL3. *J Immunol.* (2008) 180:3969–79. doi: 10.4049/jimmunol.180.6.3969
50. IBGE. *Censo 2010.* Rio de Janeiro: Atlas censo demografico (2013).
51. Braun-Prado K, Vieira Mion AL, Farah Pereira N, Culpi L, Petzl-Erler ML. HLA class I polymorphism, as characterised by PCR-SSOP, in a Brazilian exogamic population. *Tissue Antigens.* (2000) 56:417–27. doi: 10.1034/j.1399-0039.2000.560504.x
52. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA.* (1977) 74:5463–7. doi: 10.1073/pnas.74.12.5463
53. Hilton HG, Norman PJ, Nemat-Gorgani N, Goyos A, Hollenbach JA, Henn BM, et al. Loss and gain of natural killer cell receptor function in an African hunter-gatherer population. *PLoS Genet.* (2015) 11:e1005439. doi: 10.1371/journal.pgen.1005439
54. Augusto DG, Piovezan BZ, Tsuneto LT, Callegari-Jacques SM, Petzl-Erler ML. KIR gene content in Amerindians indicates influence of demographic factors. *PLoS ONE.* (2013) 8:e56755. doi: 10.1371/journal.pone.0056755
55. Kulkarni S, Martin MP, Carrington M. KIR genotyping by multiplex PCR-SSP. *Methods Mol Biol.* (2010) 612:365–375. doi: 10.1007/978-1-60761-362-6_25
56. Bunce M, O'Neill CM, Barnardo MCNM, Krausa P, Browning MJ, Morris PJ, et al. Phototyping: comprehensive DNA typing for HLA-A, B, C, DRB1, DRB3, DRB4, DRB5 & DQB1 by PCR with 144 primer mixes utilizing sequence-specific primers (PCR-SSP). *Tissue Antigens.* (1995) 46:355–67. doi: 10.1111/j.1399-0039.1995.tb03127.x
57. Vilches C, Castaño J, Gómez-Lozano N, Estefanía E. Facilitation of KIR genotyping by a PCR-SSP method that amplifies short DNA fragments. *Tissue Antigens.* (2007) 70:415–22. doi: 10.1111/j.1399-0039.2007.00923.x
58. Gross J, Ligges U. *Nortest: Tests for Normality.* (2015). Available online at: https://cran.r-project.org/package=nortest (accessed May 20, 2020).
59. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing (2019). Available online at: https://www.r-project.org/ (accessed July 06, 2020).
60. Dinno A. *dunn. test: Dunn's test of multiple comparisons using rank sums.* (2017). Available online at: https://cran.r-project.org/package=dunn.test (accessed July 06, 2020).
61. Jr FEH, Dupont C. *Hmisc: Harrell Miscellaneous.* (2019). Available online at: https://cran.r-project.org/package=Hmisc (accessed July 06, 2020).
62. Warnes G, Gorjanc G, Leisch F, Man M. *Genetics: population genetics.* (2019). Available online at: https://cran.r-project.org/package=genetics (accessed May 20, 2020).
63. Shin J-H, Blay S, Graham J, McNeney B. LDheatmap : an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Softw.* (2006) 16. doi: 10.18637/jss.v016.c03
64. Eklund A. *Beeswarm: The Bee Swarm Plot, An Alternative To Stripchart.* R package version 0.2.0. (2015). Available online at: http://cran.r-project.org/package=beeswarm (accessed July 06, 2020).
65. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature.* (2002) 419:832–7. doi: 10.1038/nature01140
66. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol Ecol Resour.* (2017) 17:78–90. doi: 10.1111/1755-0998.12634
67. Sherry ST, Ward M, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP : the NCBI database of genetic variation. *Nucleic Acids Res.* (2001) 29:308–11. doi: 10.1093/nar/29.1.308

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

REVIEW
published: 11 September 2020
doi: 10.3389/fimmu.2020.582804

# The Genetic Mechanisms Driving Diversification of the *KIR* Gene Cluster in Primates

***Jesse Bruijnesteijn*[1]*, *Natasja G. de Groot*[1] and *Ronald E. Bontrop*[1,2]**

*[1] Comparative Genetics and Refinement, Biomedical Primate Research Centre, Rijswijk, Netherlands, [2] Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands*

The activity and function of natural killer (NK) cells are modulated through the interactions of multiple receptor families, of which some recognize MHC class I molecules. The high level of *MHC class I* polymorphism requires their ligands either to interact with conserved epitopes, as is utilized by the NKG2A receptor family, or to co-evolve with the MHC class I allelic variation, which task is taken up by the killer cell immunoglobulin-like receptor (KIR) family. Multiple molecular mechanisms are responsible for the diversification of the *KIR* gene system, and include abundant chromosomal recombination, high mutation rates, alternative splicing, and variegated expression. The combination of these genetic mechanisms generates a compound array of diversity as is reflected by the contraction and expansion of *KIR* haplotypes, frequent birth of fusion genes, allelic polymorphism, structurally distinct isoforms, and variegated expression, which is in contrast to the mainly allelic nature of MHC class I polymorphism in humans. A comparison of the thoroughly studied human and macaque *KIR* gene repertoires demonstrates a similar evolutionarily conserved toolbox, through which selective forces drove and maintained the diversified nature of the *KIR* gene cluster. This hypothesis is further supported by the comparative genetics of *KIR* haplotypes and genes in other primate species. The complex nature of the *KIR* gene system has an impact upon the education, activity, and function of NK cells in coherence with an individual's MHC class I repertoire and pathogenic encounters. Although selection operates on an individual, the continuous diversification of the *KIR* gene system in primates might protect populations against evolving pathogens.

**Keywords: Killer cell immunoglobin-like receptor, KIR, NK cell, NK cell education, human, macaque, non-human primates**

OPEN ACCESS

***Edited by:***
*Jim Kaufman,*
*University of Cambridge,*
*United Kingdom*

***Reviewed by:***
*Paul J. Norman,*
*University of Colorado Denver,*
*United States*
*Francesco Colucci,*
*University of Cambridge,*
*United Kingdom*

****Correspondence:***
*Jesse Bruijnesteijn*
*bruijnesteijn@bprc.nl*

***Specialty section:***
*This article was submitted to*
*Alloimmunity and Transplantation,*
*a section of the journal*
*Frontiers in Immunology*

***Received:*** *13 July 2020*
***Accepted:*** *18 August 2020*
***Published:*** *11 September 2020*

***Citation:***
*Bruijnesteijn J, de Groot NG and Bontrop RE (2020) The Genetic Mechanisms Driving Diversification of the KIR Gene Cluster in Primates. Front. Immunol. 11:582804.*
*doi: 10.3389/fimmu.2020.582804*

## INTRODUCTION

The innate and adaptive arms of the immune system are interconnected, and feature several effector functions that provide efficient and specific protection against infection and tumor formation. Major components of the adaptive arm comprise T and B lymphocytes characterized by rearranging antigen receptors, which exert cytotoxic and humoral immunity, respectively. The cytotoxicity mediated by T lymphocytes highly depends on the presentation of intracellular antigen segments derived from pathogens by MHC class I molecules and subsequent clonal expansion of cells with specific receptors. A third type of lymphocytes bridge the innate and adaptive immune response, and comprises natural killer (NK) cells, which participate, for instance, in the recognition and elimination of aberrant cells that down-regulate their MHC class I expression to evade detection by T lymphocytes (1). Without prior priming or clonal expansion, inhibitory and activating receptors on the NK cell surface interact with MHC class I molecules on nucleated cells to modulate NK cell effector functions, which include the killing of target cells by the release of cytolytic proteins and the regulation of other immune cells by the secretion of cytokines (2). The genes encoding the MHC class I molecules are considered the most polymorphic genes known in vertebrates, a phenomenon that resulted from selective pressure to adapt to the rapid diversification of pathogens. This extended repertoire of *MHC class I* genes and alleles requires the NK cell receptors to co-evolve to maintain a functional relation with their ligands. The recognition of MHC class I molecules by NK cells involves two receptor families: the conserved CD94:NKG2A receptors and the highly polymorphic and diverse killer cell immunoglobulin-like receptors (KIR). Both receptor families consist of inhibitory and activating members. Their engagement with MHC class I molecules calibrates the responsiveness of NK cells through a continuous educational process, which largely controls subsequent NK cell activity (3, 4). The KIR receptors are encoded within the Leukocyte Receptor Complex (LRC) on chromosome 19q13.4, and share this genomic region with other structurally similar immune-regulators, such as the leukocyte Ig-like receptors (LILRs) and the leukocyte-associated Ig-like receptors (LAIRs; **Figure 1**) (5). Based on different *Alu* elements that can be regarded as a molecular clock, the initial expansion of the primate *KIR* gene cluster is estimated to date back to approximately 31 to 44 million years ago. This process continued, and is currently reflected by extensive gene duplications and point mutations (6). Different diversifying mechanisms in combination with evolutionary selective factors propel the complex *KIR* gene content at the individual level but also at the population and species-specific level, which all together contribute to the heterogeneity of NK cell subsets and their activity. The *KIR* gene diversification is not limited to humans. Comparative analyses that include other primate species might help in gaining a thorough understanding of the evolutionary processes that resulted in the diversification of this gene system. In the following sections, we will discuss the different genetic mechanisms that drove the evolution of the highly plastic *KIR* gene system in hominoids (humans and great apes) and Old World monkeys, and how this might influence their NK cell response.

## CO-EVOLUTION OF *MHC* AND *KIR* GENES

The complex *KIR* gene system requires a comprehensive nomenclature guideline for the different genes and allotypes in order to distinguish the corresponding receptors by their structure and signaling potential (7–9). Receptors may contain one to three Ig-like domains, which are encoded by exon 3 (D0 domain), exon 4 (D1 domain), and exon 5 (D2 domain), and are referred to as KIR1D, KIR2D, and KIR3D in the official nomenclature. Further classification defines the inhibitory or activating signaling function of the KIR receptors, which is characterized by either a long or short cytoplasmic tail, respectively, and specified with an "L" or an "S" following the domain number denotation. The long cytoplasmic tail contains one or two immune tyrosine-based inhibitory motifs (ITIMs), whereas the signal transduction of activating KIR depends on the interaction with an adaptor molecule that includes an immune tyrosine-based activating motif (ITAM) such as DAP12. Pseudogenes are indicated with a "P" (e.g., *KIR3DP*). In addition, a four-character species designation is included in front of the KIR acronym (e.g., *Mamu-KIR3DL20* in rhesus macaques; *Macaca mulatta*).

The mammalian *KIR* genes originate from two progenitor gene lineages: KIR3DX and KIR3DL. The KIR3DX lineage is represented by a single gene copy located in the center of the *LILR* gene cluster (**Figure 1**). The gene is fixed in most primate species, and its function is currently unknown (10). This lineage is, however, expanded in cattle, and encodes multiple inhibitory and a single activating functional KIR3DX receptor, which interact with an expanded repertoire of classical MHC molecules (11, 12). In contrast, the KIR3DL lineage expanded in primates and was diversified by duplications, deletions, and recombinations, which resulted in an elaborated *KIR* gene family. Based on their structure, ligand specificity, and/or phylogenetic analysis, the primate KIR receptors are divided into four lineages. Lineage I genes encode receptors with a D0-D2 domain configuration; lineage II (D0-D1-D2) is defined by the specificity for subtypes of HLA-A and -B in humans; lineage III includes receptors with D1-D2 and D0-D1-D2 domain configurations; and lineage V (D0-D1-D2) is represented by human KIR3DL3 and its orthologs. In the primate species studied, at least one *KIR* gene was discovered for each lineage, which indicates that gene duplication and diversification predates primate speciation. The subsequent lineage expansions are, however, species specific (**Table 1**).

Lineage I and V *KIR* genes have a conserved nature in all primate species examined, and comprise, respectively, *KIR2DL4* and *KIR2DL5*, and *KIR3DL3*, or a similar structure, such as *Mamu-KIR3DL20* in rhesus macaques. More extensive and species-specific expansions are reported for *KIR* genes that cluster into lineages II and III (**Table 1**), and the data suggest that this coincides with the evolution of their MHC class I ligands. Therefore, diversification of the lineage II and III *KIR* genes might be indirectly propelled by the adaption of the MHC class I molecules to pathogenic encounters. For hominoids, this section of co-evolution of KIR and MHC has been comprehensively reviewed by Wroblewski and colleagues (13). In short, the *MHC* gene content in great apes displays to a limited extent a variable number of *MHC-A*, *-B*, and *-C* genes per haplotype (**Table 2**). *MHC-C*, which originated from a duplication of an *MHC-B* gene, is fixed in all hominoids except for orangutans, where it is present on about half of the haplotypes (14). In addition, the epitopes recognized by the relevant KIR are differentially distributed across the different *MHC class I* genes (**Table 2**). The C1 and C2 epitopes, for example, are absent in bonobos and orangutans, respectively, whereas the A3/A11 epitope is

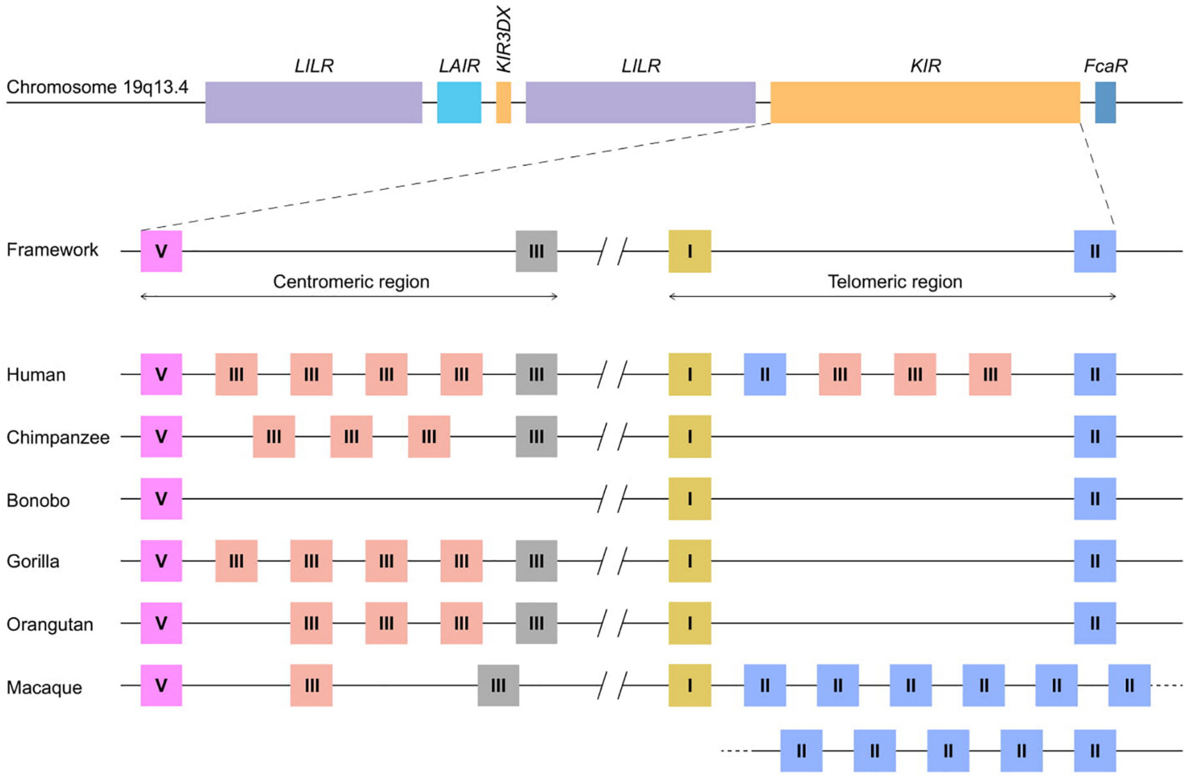


**FIGURE 1 |** KIR haplotype organizations in different primate species. A schematic overview of the Leukocyte Receptor Complex (LRC) on chromosome 19q13.4 and KIR haplotype organizations in different primate species. A fixed copy of the *KIR3DX* gene is located within the primate *LILR* gene cluster, whereas the expanded *KIR* gene cluster is flanked by the *LILR* and *FcaR* genes. The expansion involved four different *KIR* gene lineages – I, II, III, and V – the members of which are indicated as yellow, blue, red/gray, and pink boxes, respectively. The gray lineage III boxes represent pseudogenes. In most hominoids, KIR haplotype organizations follow a standard framework, in which the centromeric and telomeric regions are bordered by genes from lineages V and III, and lineages I and II, respectively. A relatively large non-coding segment separates the centromeric and telomeric haplotype sections.

only defined on HLA-A molecules. The hominoid MHC class I evolution is accompanied by the reduction and refinement of KIR specific for MHC-A and -B, which is reflected in their limited number of lineage II KIR receptors, whereas the emergence and fixation of MHC-C in humans, chimpanzees, and gorillas drove the expansion and specialization of lineage III KIR (**Table 1**) (13).

Old World monkeys, including macaques, lack an MHC-C ortholog, but instead display extensive copy number variation regarding polymorphic *MHC-A* and *-B* genes, as opposed to the fixed number of *MHC class I* genes in hominoids (**Table 2**) (15–18). The expression level of the different MHC-A and -B molecules, however, varies considerably in macaques. It is generally accepted that per haplotype at least a single *MHC-A* and 1 to 3 *MHC-B* genes are characterized by high transcription, and are referred to as "majors," whereas the other *MHC class I* genes have lower transcription levels ("minors"), or may be pseudogenes. The differential transcription suggests a more classical function for the major MHC molecules, such as antigen presentation, whereas the minors might exert more specialized functions (19, 20). Only a few interactions of macaque MHC and KIR are documented, and, so far, all interactions involved lineage II KIR that recognize Bw4 and Bw6 epitopes on MHC-A and -B allotypes (**Table 2**) (21–25). This putative lineage II specificity for the copious macaque MHC class I repertoire coincides with an extensive ligand expansion, and, thus far, 54 and 56 different lineage II *KIR* genes have been documented for rhesus and cynomolgus macaques, respectively (**Table 1**) (7). Like the majors and minors for the MHC system, the *KIR* genes, may display differential expression levels, which are modulated by sequence polymorphisms and by an individual's *MHC class I* repertoire (26–28). Lineage III *KIR* genes, which encode ligands for MHC-C in hominoids and were subject to expansion, are represented in macaques by a single gene and encodes a receptor with only the D1 extracellular domain (KIR1D). Its presence on 22% and 82% of the rhesus and cynomolgus macaque KIR haplotypes, respectively, suggests a balancing selection for this structurally modified receptor, which might execute a function other than conventional MHC recognition (29).

The maximal expression of six distinct *MHC class I* genes in most hominoids and the specialization of MHC-C as ligand for lineage III KIR is in line with their modest *KIR* gene expansion (**Tables 1**, **2**). Macaques may harbor over 20 distinct *MHC class I* genes in one individual, of which only a few are dominantly expressed and considered to be majors. The expanded MHC repertoire in macaques probably propelled the extensive expansion and differential expression of their lineage

**TABLE 1** | The number of KIR genes defined per primate species indicated per lineage.

| | | Lineage I | | | Lineage II | | | Lineage III | | | Lineage V | | | Total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Inhibi-tory | Activa-ting* | Pseudo-gene | Inhibi-tory | Activa-ting | Pseudo-gene | Inhibi-tory | Activa-ting | Pseudo-gene | Inhibi-tory | Activa-ting | Pseudo-gene | Inhibi-tory | Activa-ting | Pseudo-gene | Total |
| Human | *Hosa* | 2 | 1 | 0 | 2 | 1 | 0 | 3 | 5 | 2 | 1 | 0 | 0 | 8 | 7 | 2 | 17 |
| Chimpanzee | *Patr* | 1 | 1 | 0 | 1 | 0 | 0 | 6 | 3 | 0 | 1 | 0 | 0 | 9 | 4 | 0 | 13 |
| Bonobo | *Papa* | 1 | 1 | 0 | 3 | 1 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 7 | 2 | 0 | 9 |
| Gorilla | *Gogo* | 1 | 1 | 0 | 1 | 0 | 0 | 5 | 1 | 0 | 1 | 0 | 0 | 8 | 2 | 0 | 10 |
| Bornean orangutan | *Popy* | 1 | 1 | 0 | 1 | 1 | 0 | 2 | 3 | 0 | 1 | 0 | 0 | 5 | 5 | 0 | 10 |
| Sumatran orangutan | *Poab* | 1 | 1 | 0 | 1 | 1 | 0 | 3 | 3 | 0 | 1 | 0 | 0 | 6 | 5 | 0 | 11 |
| Rhesus macaque | *Mamu* | 0 | 1 | 0 | 31 | 23 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 33 | 24 | 1 | 58 |
| Cynomolgus macaque | *Mafa* | 0 | 1 | 0 | 26 | 30 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 28 | 31 | 1 | 60 |

**KIR2DL4 is considered an activating KIR gene.*

II KIR. The balanced expansion of the *MHC* and *KIR* gene systems in primates indicates co-evolution in order to maintain a functional relation.

## TRANSPOSABLE ELEMENTS FACILITATE CHROMOSOMAL RECOMBINATION

One of the mechanisms responsible for the extensive *KIR* gene diversification in macaques, and to a lesser extent in hominoids, involves chromosomal rearrangements that are accompanied by deletions and recombination, which may result in the generation of fusion genes (**Figure 2A**). This type of gene formation may shuffle the binding and signaling domains of different KIR receptors, thereby functionally altering the response potential of KIR family members. The dense head-to-tail arrangement of the *KIR* genes is likely to facilitate at least in part the chromosomal instability of this gene cluster. A KIR haplotype spans approximately 150 to 350 kb, depending on the number of genes present. Most *KIR* genes are separated by only 2.5 kb, as opposed to the wider haplotype configurations of more stable and less expanded gene families, such as the *LILR* gene cluster (6, 30–33). In addition, the presence of transposable elements, including Alu and LINE elements, in the intergenic and intragenic KIR sequences is another factor that further promotes genetic instability (6, 34–36). These repetitive elements are present in all primate *KIR* genes, although with species-specific variation, and drive recombination and genetic deletions (35, 37–39). For the few completely sequenced fusion *KIR* genes in humans, the chromosomal breakpoints indeed map in the intragenic transposable elements. This supports the idea that the abundant presence of transposons in the *KIR* cluster facilitates chromosome fragility, which is reflected by genetic expansion and contraction, and the formation of fusion genes (34, 40, 41). A considerable number of human fusion *KIR* genes were generated by reshuffling that involved segments of pseudogenes (34). The conservation of two pseudogenes in the human KIR repertoire, *KIR2DP1* and *KIR3DP1*, might be explained by their role in promoting recombination events. The human KIR haplotypes that include an apparent fusion gene are represented by relatively low frequencies (42–45). Positive selection of fusion entities might, however, increase their frequencies in certain populations (45). Ancient recombination events and subsequent selection might have contributed substantially to the current human KIR repertoire, but the modest expansion of the human *KIR* genes nowadays indicates limited recent recombination events. In contrast, an excessive number of recombination events are recorded in rhesus and cynomolgus macaques, with the presence of at least one fusion *KIR* gene on 42% and 49% of their haplotypes, respectively (29, 43). The abundant presence of fusion genes indicates that in these species the reshuffling of *KIR* gene segments is an ongoing process that expands the macaque KIR repertoire. Although information on the non-coding regions in the macaque *KIR* cluster is limited at present, the chromosome instability and consequential recombinations in concert with selection are likely to have driven the extensive

**TABLE 2 |** The expansion of MHC class I genes in different primate species.

| | MHC-A | | MHC-B | | MHC-C | |
|---|---|---|---|---|---|---|
| | **# genes** | **KIR-epitopes** | **# genes** | **KIR-epitopes** | **# genes** | **KIR-epitopes** |
| Human | 1 | A3/A11, Bw4 | 1 | Bw4, C1 | 1 | C1, C2 |
| Chimpanzee | 1 | – | 1 | Bw4, C1 | 1 | C1, C2 |
| Bonobo | 1 | – | 1 | Bw4, C1 | 1 | C2 |
| Gorilla | 1 (2*) | Bw4 | 1–2 | Bw4, C1 | 1 | C1, C2 |
| Orangutan | 1 | – | 2–4 | Bw4, C1 | 0–1 | C1 |
| Macaque | 1–3 | Bw4, Bw6 | 1–3 (<19) | Bw4, Bw6 | – | – |

*Indicated are the number of genes present on a single chromosome and the KIR-recognizing epitopes that may be encoded by allotypes. The frequencies of the different epitopes vary per gene and species. In macaques, on average 1–3 MHC-B genes are highly transcribed (majors), whereas the total number of genes on a single MHC haplotype can reach up to 19 copies, including low transcribed genes (minors) as well as pseudogenes (15–20). *Gorilla's have an additional MHC-A related gene, named Gogo-Oko.*

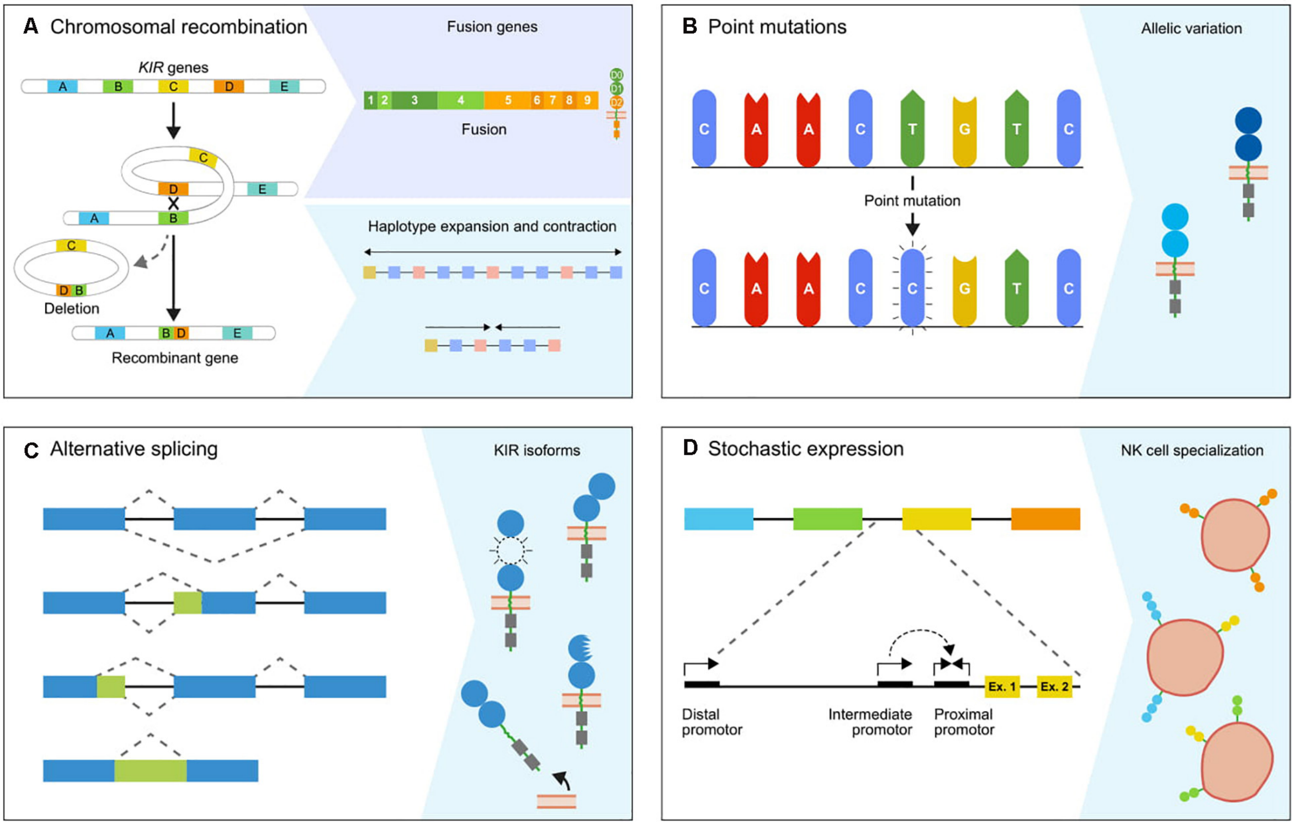


**FIGURE 2 |** The genetic mechanisms propelling diversification. The primate KIR cluster diverged as a result of multiple molecular processes, which together modulate the *KIR* gene content and expression status. **(A)** The expansion and contraction of KIR haplotypes is mediated by chromosomal recombinations, which can introduce or remove one or multiple *KIR* genes. Occasionally, a recombination event is accompanied by the generation of a fusion gene, which functionally and structurally expands the gene repertoire. **(B)** The *KIR* genes are further diversified by point mutations in coding and non-coding regions, which generate alleles that encode receptors with different structures, localization, function, and expression. **(C)** Alternative splicing is another mechanism that has a similar impact on the function and structure of receptors. The blue and green boxes indicate exons and introns. The isoforms are generated by different splice events, which involve alternative splice sites and exon skipping. **(D)** The differential expression of subsets of KIR receptors on different NK cell clones forms another level of variation that is mediated and maintained by sequence variability in the complex promoter regions and epigenetic modifications. A conjunction of the proximal, intermediate, and distal promoters is required to induce KIR expression.

expansion of lineage II *KIR* genes. This fast mode of evolution is further reflected in the relatively low number of orthologs that are shared between the closely related rhesus and cynomolgus macaques and their populations (29).

In all hominoids and Old World monkeys, the 5' section of the *KIR* gene cluster is occupied by *KIR3DL3* or similar structures, which are considered framework genes and might carry out essential functions. The structure and evolutionary pathway of these lineage V *KIR* genes is a complex outcome of multiple recombination events (46). Additional chromosomal rearrangements in rhesus macaques involved the exchange of the cytoplasmic tail of *KIR3DL20* with the tails of *KIR2DL04* (lineage I) and *KIR1D* (lineage III). These recombination events are not conserved in macaque populations, which implies the relatively recent formation of novel gene entities propelled by ancient recombination hotspots (29).

Chromosomal recombination events generate genetic variability in the *KIR* gene cluster by the formation of fusion genes. Subsequent selection of these novel genes might supply an adaptive and protective strategy in the arms race with rapidly evolving pathogens.

## KIR HAPLOTYPE DIVERSITY IN PRIMATE SPECIES

Chromosomal rearrangements not only generate novel *KIR* gene entities by recombination but also diversify the haplotype gene content by insertions and deletions of genes (**Figure 2A**). In general, hominoid KIR haplotypes consist of two genomic regions that are bordered by four framework genes (**Figure 1**). The proximal half of the haplotype is termed the centromeric region and is defined by *KIR3DL3* to *KIR3DP1/KIRDP*, whereas the distal part, or telomeric region, ranges from *KIR2DL4* to *KIR3DL2/KIR3DL1*. Within these sections, *KIR* genes of different lineages expanded and contracted during hominoid speciation. In humans, the expansion involved lineage III *KIR* genes in their centromeric and telomeric regions, whereas expansion in chimpanzees, gorillas, and orangutans expansion took place in the same lineage in the centromeric region only (**Figure 1**). The human haplotype content ranges from 7 to 12 *KIR* genes, whereas the number in chimpanzee and orangutan haplotypes stretches from 5 to 11 and 5 to 10 functional *KIR* genes, respectively. In contrast to other hominoids, bonobos are characterized by a contraction of their KIR region, with only 3–7 *KIR* genes expressed on a haplotype. The shortest bonobo KIR haplotype consists of only the framework genes (47). The contracted bonobo KIR cluster coincides with a reduced nucleotide variation in their MHC class I repertoire, which might be caused by a bottleneck or pathogen-driven selective sweep after divergence from the chimpanzee's lineage (48–51). In contrast, a highly variable *KIR* haplotype content is encountered in the macaque, with 4 to 17 functional *KIR* genes that mainly map to the telomeric region (**Figure 1**). The haplotype framework in macaques is less fixed than in hominoids, with only *KIR3DL20* expressed on all haplotypes, whereas *KIR2DL04* is present on 70% of the rhesus macaque haplotypes. A gene orthologous to hominoid *KIR3DL2*/*KIR3DL1* that usually marks the telomeric region is absent.

The diversified *KIR* haplotypes in hominoids and Old World monkeys stem from a primordial configuration, for which a model has been proposed by Guethlein and colleagues (35). This model describes abundant duplications and recombination events that eventually formed a conserved haplotype framework in hominoids. The previously mentioned transposable elements are likely propagating these chromosomal rearrangements that continue to mediate the diversification of haplotype configurations. One major hotspot for recombination is mapping in between the centromeric and telomeric regions, which facilitates the reorganization of the different haplotype regions. In addition, KIR haplotypes also display the insertion and deletion of one or multiple *KIR* genes propelled by unequal crossing-over, which is occasionally accompanied by the formation of a fusion gene (29, 40, 42–44). In humans, these contractions and expansions, which are mediated by double-stranded breaks at intragenic and intergenic repetitive elements, resulted in haplotypes that expressed 3 to 15 *KIR* genes (40). The short haplotypes do not express all framework genes. For instance, the deletion of *KIR2DL4* is commonly observed on genotypes defined across different populations (52, 53). Approximately 7% of the human *KIR* haplotypes are showing indications for contraction and expansion (42). Although the number of completely defined *KIR* haplotypes in other hominoids is low, several rare *KIR* configurations in chimpanzees and orangutans illustrate genetic footprints for insertion and deletion events, which is also occasionally accompanied by the formation of a fusion gene (54, 55). In macaques, only two completely sequenced haplotypes are available at present, whereas an abundant number of haplotypes are deduced at the transcription level by segregation studies (26, 29, 31, 43, 56, 57). The presence of multiple highly similar allotypes, encoded by highly similar *KIR* genes, on a single haplotype indicates an expansion by the insertion of one or more genes. Such events were recorded for 47% and 26% of the rhesus and cynomolgus macaque haplotypes, respectively (29). In contrast, a minimal *KIR* gene content and the presence of a fusion gene often are indicative of a haplotype contraction. An example of a prominent haplotype reduction in rhesus macaques involved the deletion of the complete centromeric region by an intragenic recombination of *KIR3DL20* and *KIR2DL04* (29). The variable haplotype content and the relatively high number of fusion genes indicate extensive recombination as a mechanism to diversify the macaque *KIR* gene system in a still ongoing process. This phenomenon is observed to a lesser extent for the *KIR* haplotypes in hominoids, where the process seems to have relaxed.

## THE *KIR* GENE ALLELIC REPERTOIRE IS EXPANDED BY POINT MUTATIONS

Another level of variation is displayed by allelic polymorphisms, which is explained to a large extent by the occurrence of single nucleotide polymorphisms (SNP; **Figure 2B**). These nucleotide variations have a wide-ranging impact, and may modulate the expression level at the cell surface, ligand specificity, interaction strength, and localization of the KIR receptor. Single nucleotide variations in the extracellular D0 and D1 domains of human *KIR2DL2**004* and *KIR3DL1**004*, for example, retain the receptors within the cell, which might be caused by misfolding (58, 59). Polymorphisms in *KIR2DL3* alleles affect the avidity of the receptor to bind their HLA-C ligands. The low-avidity *KIR2DL3**001* and the high-avidity *KIR2DL3**005* only differ at three nucleotides in their D1 domain, which alters the orientation of their extracellular domains and thereby their binding strength (60). Although most KIR disease association studies determine the gene content by the presence and absence of *KIR* gene sections, and thereby lack allele-level resolution, several studies demonstrated that the functional differences of *KIR* alleles might also impact health and disease. For example, two *KIR2DL1* alleles in the African KhoeSan population evolved by single nucleotide mutations and are associated with a reduced risk for pregnancy disorders (61). Other associations demonstrated that the highly expressed *KIR3DL1* alleles are more protective against disease progression in HIV-infected individuals than lower expressed allotypes, except for the intracellularly retained *KIR3DL1**004*, which is low in expression but highly protective (62–64).

A total of 1110 human *KIR* alleles are cataloged in the Immuno Polymorphism Database (IPD-KIR, release 2.9.0), whereas the number of reported alleles for different non-human primate species ranges from 521 *KIR* alleles in rhesus macaques to 5 *KIR* alleles in Bornean orangutans (IPD-NHKIR, release 1.2.0.0). These allele numbers may give a distorted view of the actual levels of polymorphism within a species due to the differential number of individuals studied. The high level of allelic polymorphism appears to be at least comparable in humans and macaques. The thoroughly documented allelic polymorphism in humans and macaques reveals a varying number of alleles per *KIR* gene, with most nucleotide variation exhibited by the framework genes (7, 29, 44). In addition, a high number of alleles were reported for certain *KIR* genes located on the telomeric haplotype region in humans (*KIR3DL1*, *KIR2DS4*) and the highly frequent inhibitory *KIR* genes in macaques (*KIR3DL01*, *KIR3DL07*). An expansion of the allele numbers for the frequently expressed *KIR* genes might indicate a continuous role in co-evolution with particular pathogens. The less common *KIR* genes, which include mostly activating KIR, vary in gene content rather than allelic polymorphism and therefore seem to execute more specialized functions and/or might be involved in the recognition of conserved ligands and peptides (7, 29, 44).

For humans, *KIR* alleles are also distinguished by SNPs in their introns (IPD-KIR, release 2.9.0) (65), which might impact, for instance, the expression level and post-transcriptional splicing. A total of 353 human *KIR* alleles can only be distinguished from the reference gene based on intronic variations (IPD-KIR, release 2.9.0), and this number is likely to be underestimated (65). Sequence data on the non-coding *KIR* gene regions are lacking for non-human primate species, but a similar extent of intronic variations might be feasible and may impact their receptor functionality. However, there are no disease or health associations reported for intronic polymorphisms within the *KIR* genes, but abundant pathological conditions are described for intronic variations in many other genes mapping elsewhere in the genome (66). For example, a SNP in the human *CYP2D6* gene is linked to a decreased expression of the functional transcript and correlates with a lower metabolic activity (67). For *HLA-DP*, a single nucleotide variation in the 3′ UTR modulates the expression level of different allotypes, which impacts the susceptibility to chronic hepatitis B virus infection (68).

Allele variation is mainly generated by synonymous and non-synonymous point mutations, and only the latter ones will impact the composition of the gene products. In sharp contrast to MHC class I polymorphisms, the allelic nucleotide variations of the *KIR* genes are evenly distributed over the coding regions. The high concentration of CpG islands located in the *KIR* gene cluster might contribute to an elevated mutation rate, as these islands are in general more prone to promote nucleotide transitions (69–71). In addition, chromosomal rearrangements are known as mutagenic events (69, 72–74). In particular, the regions that surround genomic insertions and deletions display an increased mutation rate, which might be induced by error-prone DNA replication (69, 75–77). The abundant recombination that is accompanied by insertions and deletions in the primate KIR cluster is likely to contribute to the extensive allelic KIR variation. Within two and three generations of human and macaque families studied, the birth of novel KIR alleles is described, which might further substantiate the rapid mutation rate in this gene cluster (29, 78). To our knowledge, such an event has not been recorded for the highly polymorphic *MHC class I* genes.

The variation involving *KIR* genes at the allele level impacts the interactions with their highly polymorphic MHC class I ligands, and demonstrate that point mutations contribute to a diversified *KIR* gene system. The general lack of allele level characterization in the clinic might limit the number of associations reported for *KIR* allele heterogeneity and their functional and disease-related effects. Even intronic variations might impact the KIR receptor expression and function. These few associations highlight the need to further characterize the *KIR* gene content of humans and other primate species at an allele level resolution.

## ALTERNATIVE SPLICING AS A MECHANISM FOR STRUCTURAL DIVERSIFICATION

The complexity of the primate *KIR* gene cluster is further extended by alternative splicing (**Figure 2C**) (79–83). This post-transcriptional mechanism can generate multiple messenger RNA (mRNA) transcripts from a single gene, which are translated into different receptor isoforms. Constitutive splicing excludes the intronic sequences from the precursor mRNA (pre-mRNA) and ligates the coding exons. Alternative splicing deviates from this pattern by the use of alternative splice sites, the skipping of exons, and the retention of introns (**Figure 2C**) (84). The alternative splice events for human and macaque KIR transcripts are well documented, and demonstrated that both in- and out-of-frame transcripts are generated (79–83). The out-of-frame transcripts often have an early stop codon, and this results in early truncation of the transcript. Even though these out-of-frame transcripts appear as a redundant side effect of alternative splicing, it might reflect a regulatory pathway to rapidly down-regulate receptor expression. The functional impact of the in-frame generated KIR isoforms may be diverse. The skipping of exons generates transcripts that encode modified KIR isoforms, which lack one or two extracellular domains, the stem region, or the transmembrane region. These KIR isoforms probably exhibit differential binding properties or are secreted as soluble receptors (**Figure 2C**) (85). In-frame splice events that involve alternative splice sites might insert a partial intronic sequence into the transcript or delete a part of a coding exon. Although the functional and structural consequences of these KIR isoforms are harder to predict, they are likely to modify the receptor expression level, cellular localization, and ligand interactions.

Several splice events were frequently recorded or were defined for multiple *KIR* genes, and implicate the existence of conserved splice events that generate structurally and functionally distinct isoforms. For example, exon 4 (coding for the D1 domain) is frequently skipped from KIR3DL20 transcripts in macaques, thereby generating transcripts that encode both the complete receptor and receptors with a D0-D2 domain configuration (43, 57). This macaque isoform is termed *KIR2DL05*, as it displays an 89.5% sequence similarity with human *KIR2DL5*. Moreover, it demonstrates that alternative splicing expands the macaque KIR repertoire by generating a second two-domain structure (KIR2DL) additional to KIR2DL04. The most frequent KIR splice event in humans involved the skipping of exon 6, which encodes the stem region. Other frequent events included the skipping of exon 5 (D2 domain) and partial deletions in exons 4 and 5. These events result in isoforms that are likely to display altered binding properties, but their exact activity and localization remains elusive. Another common splice event in humans might function as a regulatory switch for expression of the 9A and 10A *KIR2DL4* alleles by restoring or disrupting the open reading frame (ORF) (79). Less frequent alternative splicing events were often found to be gene specific, and were mainly out-of-frame events that encoded for truncated receptors. Except for most exon skipping events, only a single splice event was shared between humans and macaques. This event involved a partial deletion of exon 3 (D0 domain) mediated by an alternative 5' splice site (79). Data on the alternative splicing in other hominoids are lacking, but a similar extent of alternative splicing is likely to diversify their KIR receptors and repertoire.

The splicing of pre-mRNA not only facilitates diversification of the KIR repertoire, but might also compensate for genomic alterations that result in out-of-frame transcripts. The expression of human and macaque lineage III *KIR* genes, for example, requires the constitutive skipping of exon 3 to maintain an ORF. This exon contains a deletion of 5 bp at the genomic DNA level, which would shift the reading frame that introduces an early stop codon (79, 86). The constitutive skipping of exon 3 suggests that the expanded repertoire of human KIR2D receptors evolved from a *KIR3D* gene. The absence of a conserved 33 bp sequence in intron 2 of all human and macaque lineage III *KIR* genes might relate to the constitutive exon skipping by, for example, disrupting the spliceosome recognition site (79).

The extensive levels of alternative splicing observed in humans and macaques defines another layer of complexity for the *KIR* gene cluster. This diversifying mechanism generates structurally and functionally distinct receptor isoforms, and might be involved in the regulation of receptor expression levels. Although not all isoforms might be functional, the frequency and consistency of several alternative splicing events suggest that alternative splicing is a rapid mechanism to diversify the KIR content in hominoids and Old World monkeys.

## DIFFERENTIAL NK CELL POPULATIONS DUE TO VARIEGATED *KIR* GENE EXPRESSION

*KIR* gene plasticity is further reflected by the stochastic expression of a subset of *KIR* genes from the total gene repertoire in individual NK cells (**Figure 2D**). This selective transcriptional activation generates specialized NK cell populations, which express different numbers and combinations of *KIR* genes (87, 88). The stochastic KIR expression is activated during NK cell maturation, and the transcriptional pattern is maintained by the methylation of silenced *KIR* genes (28, 89). The different KIR receptor combinations are generated largely at random, but might be shaped by the individual *KIR* gene frequencies and the MHC class I repertoire. Therefore, *KIR* genes that are present on both chromosomes in heterozygous individuals, or genes that are present as two or more allotypes on a single haplotype (e.g., by duplication or gene insertion), could be expressed in a mono- and multi-allelic manner. This may generate NK cell subsets that transcribe two or more allelic copies of a certain *KIR* gene (28). Divergent expression patterns are documented for human *KIR2DL4*, which is expressed in all NK cells, and for *KIR3DL3*, which is expressed at low levels (90, 91).

The molecular regulation of *KIR* gene expression is well studied in humans, and involves multiple promoter regions in the intergenic sequences that control gene demethylation and transcription (27, 90–96). The proximal promoter is located directly in front of the first exon of a *KIR* gene and functions as a probabilistic switch (**Figure 2D**). Bi-directional transcription of this promoter generates forward and reverse transcripts that correlate with the activation and suppression of *KIR* gene transcription, respectively. Forward transcripts of a distal promoter are associated with activation of the proximal promoter region and appear to be required for eventual *KIR* gene expression. A third promoter upstream of the proximal promoter, also denoted as the intermediate promoter, modulates the bidirectional transcription of the proximal promoter directly or indirectly by mediating correct splicing of the forward proximal promoter transcripts (27, 94). In all human *KIR* genes, the promoter regions are highly conserved, with 91–99.6% sequence similarity. Exceptions are found for the promoters of *KIR2DL4* and *KIR3DL3*, which substantiates their diverged expression profile (94). Three types of promoter regions are defined for human *KIR2DL5*, which display considerable differences in their nucleotide sequence and transcription factor binding sites. Types I and III control variegated expression, whereas transcripts of *KIR2DL5* alleles that exhibit the type II promoter are undetectable (97, 98). These type II promoters are probably inactivated by a SNP in their Runt-related transcription factor (RUNX) transcription binding site, which is an important motif in the regulation of gene expression, and is generally conserved in all *KIR* genes (98).

An identical SNP is identified in the proximal promoter of the pseudogene *KIR3DP1*, and might indicate that the inactive type II promoter is swapped to particular *KIR2DL5* alleles by chromosomal recombination (98–100). Within the *KIR* promoter regions, multiple other transcription factor binding sites are identified, which can vary per *KIR* gene and thereby contribute to differential gene expression. Allelic variations of the different transcription factor binding sites modulate the expression levels of *KIR* alleles (27, 92). For example, a *KIR2DL1* allele displayed low expression, which was associated with three SNPs in the distal promoter that generated a binding site for the Zinc finger E-box-binding homeobox 1 (ZEB1) protein (27). This transcription factor is associated with the down-regulation of IL2 expression, and might have a similar impact on the expression of this specific *KIR2DL1* allele. Just like the variation in the *KIR* gene introns, the nucleotide polymorphisms in the promoter regions are grossly undervalued, despite the direct impact on the expression of KIR alleles.

The variegated expression pattern of the *KIR* genes defines NK cell subsets, of which several are tissue resident. These NK cell populations might execute specialized functions in particular tissues that could be mediated by specific sets of KIR receptors. For example, the KIR expression profile of NK cells that were derived from the lung, liver, and uterus deviates from the expression pattern observed in peripheral blood NK cells (101–103). Expression of KIR was also established for subsets of T cells, in particular terminally differentiated CD8 + T cells, of which 30% exhibited KIR expression (104–106). The majority of these T cells dominantly express a single inhibitory or activating *KIR* gene, which is generally distinct from the *KIR* gene expression pattern on NK cells within the same individual (104). The expression pattern of NK cells and CD8 + T cells can be erased by *in vitro* treatment with a methylation inhibitor (5-azacytidine), and thereby induce the expression of formerly silenced *KIR* genes (28, 96, 107). This demonstrates that the stochastic *KIR* gene expression is maintained by methylation in both types of lymphocytes.

The variability in the promoter regions that is mainly generated by point mutations and chromosomal recombination events contributes to the diversification of NK cell subsets by the stochastic methylation of *KIR* genes. The promoter regions and epigenetic regulation of the *KIR* gene cluster in non-human primate species are less well characterized, but their stochastic expression pattern indicates a similar genetic mechanism.

## THE DIFFERENT CHARACTERS OF DIVERSIFICATION IN THE *KIR* AND *MHC* CLUSTERS

The expansion of the primate *KIR* cluster was probably initiated by the integration of multiple retroviral elements near or in the founding *KIR* genes. Subsequent duplications were mediated by these transposable elements, and this process had an impact on the expansion of the *KIR* gene repertoire (35). These recombination events might have enhanced the mutation rate within this genomic region that generated a diverse set of *KIR* alleles, and subsequently some of these were positively selected during evolution. In the case of exons, the point mutations may affect the receptor structure, function, localization, and expression, whereas polymorphisms in the introns may enhance the level of alternative splicing by affecting existing or generating alternative splice sites. In addition, the high level of point mutations caused variation within the promoter regions, and thereby modulated the variegated expression pattern and expression level of KIR receptors. It appears that all the different molecular mechanisms are intertwined and enhanced by each other, which multiplies their diversifying impact on the primate *KIR* gene system.

The *MHC class I* gene family is considered one of the most polymorphic genomic regions in primates, but displays a different nature of diversity as compared to its KIR ligands. In hominoids, the fixed number of *MHC-A*, *-B*, and *-C* genes on a haplotype indicate low levels of recent duplications and chromosomal recombination, which is substantiated by an exceptionally low recombination rate for the *MHC class I* region (108, 109). Chromosomal rearrangements that are accompanied by the formation of an *MHC class I* fusion gene, as is determined for the *KIR* genes, is to our knowledge not known. In most hominoids, *MHC class I* polymorphism is mainly generated by point mutations in concert with a recombination of small segments. These genetic modifications are especially located in the exons encoding the peptide-binding site, and indicate a rigorous selection for a diverse array of allotypes. The functional impact is reflected in differential peptide presentation (18). Additional modification of the MHC repertoire is reflected at the transcription level by alternative splicing, which is reported for human and macaque MHC transcripts (110–114). Considering the high level of allelic polymorphism in the *HLA* genes, which may involve nucleotide substitutions that disrupt existing or generate novel alternative splice sites, one might expect abundant alternative splicing events in their transcripts. However, only a modest level of alternative splicing is demonstrated for several classical and non-classical *HLA class I* alleles, which mainly involved exon skipping that abrogated receptor surface expression (110). Specific isoforms of the non-classical *HLA-G*, however, are well known and are associated with cancer and inflammatory diseases (115–118). In contrast, alternative splicing in primate *KIR* was not limited to certain alleles, and also comprised conserved splice events that were common to multiple *KIR* genes and lineages (79). The classical MHC class I allotypes are constitutively expressed on all nucleated cells, and thereby lack a variegated expression pattern (119, 120). However, individual MHC allotypes may display a differential expression level, which is affected by sequence variation, tissue distribution, and pathogenic encounters (120, 121). In humans, the relative surface expression of HLA-A and -B is approximately ten times higher compared to HLA-C molecules (120, 122). This suggests that the *HLA-C* gene might slowly shift its main function from classical antigen presentation into the modulation of NK cell responses during infection and reproductive biology. In addition, the expression levels of different HLA-C alleles display variation, in which highly expressed allotypes correlated with a beneficial control of HIV infection (123). The differential expression pattern is also determined for the expanded MHC class I region in macaques, with only a few highly expressed MHC-A and -B allotypes (19, 124). The MHC expression levels are, however, not strictly maintained and can be modulated during infection by immune regulators such as interferon and tumor necrosis factor (TNF) (120).

The primate *KIR* and *MHC* gene families are both reflected by great complexity, and seem to co-evolve to maintain a functional relationship. The *MHC class I* diversification mainly involved allelic polymorphism in the exons encoding the peptide binding site and recombination of small segments, which is driven by the arms race with rapidly evolving pathogens. The *KIR* genes, in contrast, are diverged by haplotype expansion and contraction, random point mutations, and the generation of novel fusion genes. The expression and structural variability of the KIR receptors are further modified at the epigenetic and post-transcriptional level, whereas a similar diversification of the MHC class I molecules is limited. The conjunction of different genetic mechanisms generates an extensive plasticity for the primate *KIR* gene cluster, which seems to exceed the diversity of the polymorphic *MHC class I* genes.

## CD94:NKG2A- OR KIR-DEPENDENT EDUCATION IN DIFFERENT PRIMATE SPECIES

A comparison of the *KIR* gene system in primate species illustrates a variable degree of gene expansion, reflected in the differential expansion of gene lineages (**Figure 1**). This might be largely due to co-evolution with their diverse MHC class I repertoire. The variable extent of expansion, however, is emphasized by the number of functional genes per *KIR* haplotype and by the overall size of the *KIR* gene repertoire documented for a certain primate species. The extremes are represented by the heavily contracted KIR haplotypes in bonobos versus the widely expanded set of *KIR* genes in macaques (**Figure 1**). The flexibility to expand and contract *KIR* haplotypes and repertoires, apparently without compromising sufficient and protective immune responses, might be closely related to the nature of NK cell education in different primate species.

Natural killer cells require self-tolerance and a signal to activate, which are acquired through an educational process. NK cell education involves the recognition of self-MHC class I molecules or the presented peptides by at least one inhibitory NK cell receptor. Alternative educational pathways that are MHC-independent are reported, but their exact contribution to the acquiring of NK cell functions is elusive (125, 126). The MHC-dependent education is predominant and can be approached in two ways (**Figure 3**) (13, 127, 128). One strategy of NK cell education involves the interaction of inhibitory CD94:NKG2A NK cell receptors with the non-polymorphic MHC-E molecules, which are complexed with conserved signal peptides derived from the diversified classical MHC class I molecules (129–131). One could argue that this approach allows the immune system to scan in a crude way whether total MHC class I expression has been abrogated. In the complementary approach, however, NK cell education is established through interaction of the MHC class I molecules with polymorphic KIR receptors. This seems to reflect a more sophisticated strategy in which the immune system checks at the epitope level for a malfunctioning of MHC class I expression. KIR-dependent NK cell education is mainly conducted through the interactions of inhibitory KIR and MHC class I molecules. However, activating KIR contribute to the tuning of NK cell responsiveness by dampening NK cell activity upon MHC class I recognition (132). Currently, only for KIR2DS1 the effect on NK cell education is described. In the following sections, we mainly consider the educational impact of inhibitory KIR.

Whether the NK cells are educated by the CD94:NKG2A or KIR pathway might depend on a single nucleotide dimorphism at position 21 of the MHC class I leader sequences. Most MHC-A and -C molecules in hominoids have a methionine (−21M) residue present at this position, whereas in general this position is occupied by threonine (−21T) in MHC-B molecules. The -21M peptides strongly bind to MHC-E molecules and promote cell surface expression of MHC-E complexes (133). The presence of five or six classical MHC class I allotypes containing the -21M residue drives the NK cell education toward the more conserved MHC-E and CD94:NKG2A interactions. However, approximately 62% of human individuals display a -21T HLA-B homozygous genotype, with a variable distribution in different populations (127). In chimpanzees, -21T is near fixed in their MHC-B allotypes (13). The homozygous threonine genotype corresponds with a low MHC-E surface expression. As a consequence, human and chimpanzee NK cells are largely educated by their KIR repertoire (13, 127). In contrast, in macaque MHC-A and -B allotypes, methionine is the predominant residue at position 21 of the leader sequence, which results in an NK cell education that mostly relies on the conserved CD94:NKG2A pathway (127).

In primate species with a KIR-dependent NK cell education, one can envision that an expanded KIR repertoire may compromise NK cell activity. This might drive selection for a limited KIR expansion, as we will discuss in the next section. If this reasoning is true, the KIR-independent education of NK cells in macaques might result in an extensive expansion of their *KIR* gene system. We think that the primary function of macaque KIR is focused on the recognition and elimination of infected or malignant cells. This defense mechanism relies on the recognition of Bw4 and Bw6 epitopes, but KIR interactions are also sensitive to non-self peptides that can be presented by MHC class I molecules (134–138). A large genetic diversity of *KIR* genes provides a broader repertoire to scan all the variable MHC class I allotypes in combination with their peptides originating from pathogens. It has been proposed that up to seven distinct KIR receptors are required for successful peptide recognition (139). This optimal receptor count might even be higher when the Bw4 and Bw6 epitope specificity is considered for the different KIR allotypes. The high level of chromosomal recombination and the relatively frequent formation of fusion genes in macaques might indicate selection for a widely diversified *KIR* gene system. Considering their KIR-independent NK cell education,

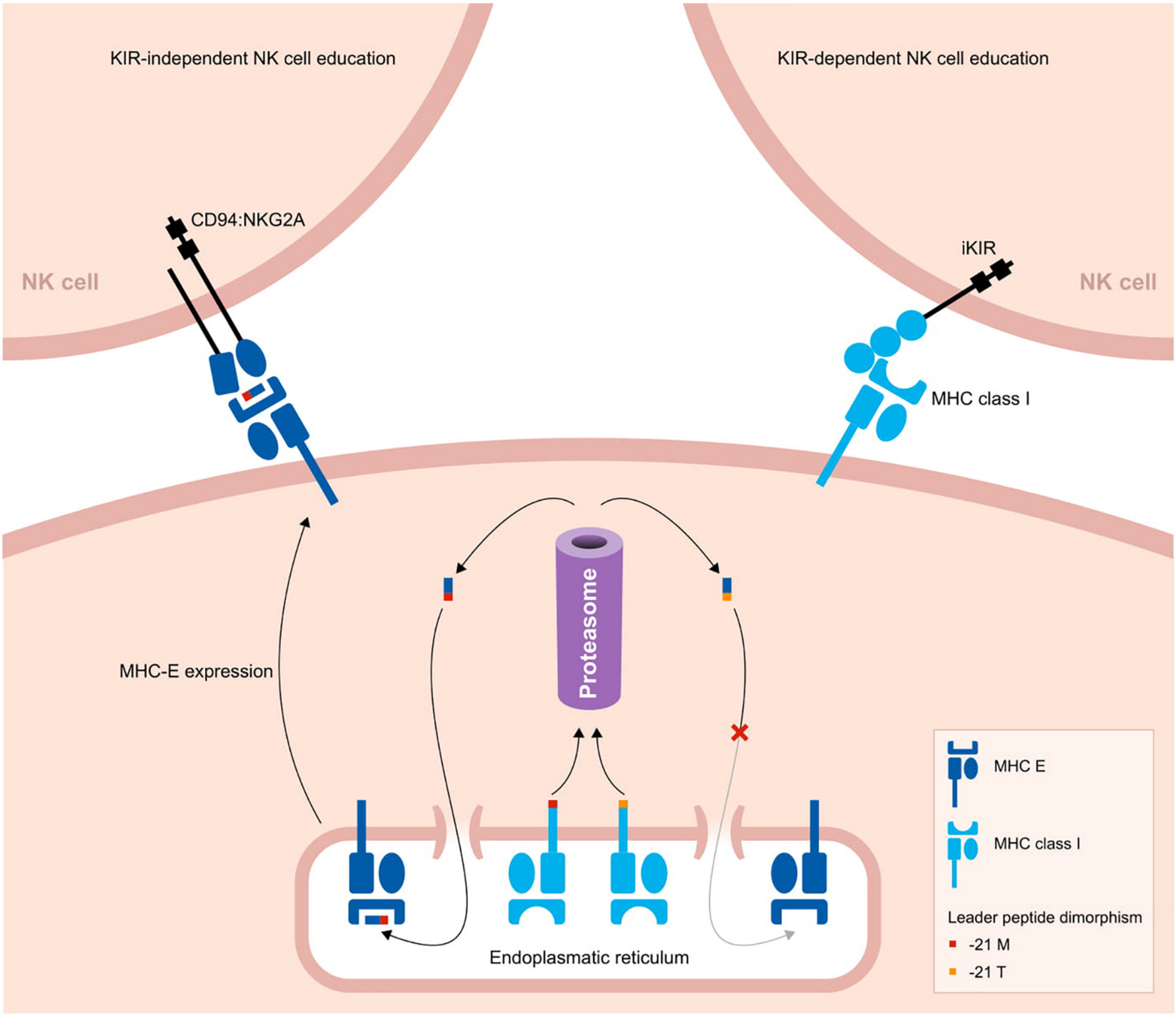


**FIGURE 3 |** Two pathways to educate NK cells in primates. A schematic overview of two strategies to educate NK cells. The leader peptide of MHC class I molecules either contains a methionine (−21M) or threonine (−21T) residue. The -21M peptides strongly bind to MHC-E molecules, and these complexes display an increased cell surface level. Through the conserved CD94:NKG2A receptors, the MHC-E complexes educate and license NK cells. In contrast, -21T leader peptides, which are predominantly present in MHC-B allotypes of humans and chimpanzees, do not interact with MHC-E molecules. Therefore, in the presence of one or more -21T MHC class I allotypes, the cell surface level of MHC-E is decreased. In this case, NK cells are educated by MHC class I molecules that interact with their KIR ligands.

KIR expansion in macaques might be exempted from potential negative selection on large *KIR* gene repertoires.

However, not all macaque *KIR* haplotypes contain a large number of genes, and they even display indications for contraction by chromosomal recombination events. The formation of novel gene entities by the shuffling of head- and tail-encoding exons is achieved by recombination events, which are coherently accompanied both by contractions and expansions of *KIR* haplotypes. There might be a trade-off between the expansion of the overall KIR repertoire in a population by generating fusion genes and the contraction of KIR haplotypes in individuals. Rapid expansion and diversification generate a highly plastic macaque *KIR* gene system that appears to be maintained by selection to militate against rapidly evolving pathogens.

## KIR HAPLOTYPE EXPANSION AND CONTRACTION: FINDING THE EQUILIBRIUM

As compared to macaques, hominoids appear to have a more limited haplotype content and overall KIR repertoire (**Figure 1** and **Table 1**). These limitations might be maintained by selective pressure on an efficient KIR-dependent NK cell education, but should be balanced with protection against infections. This

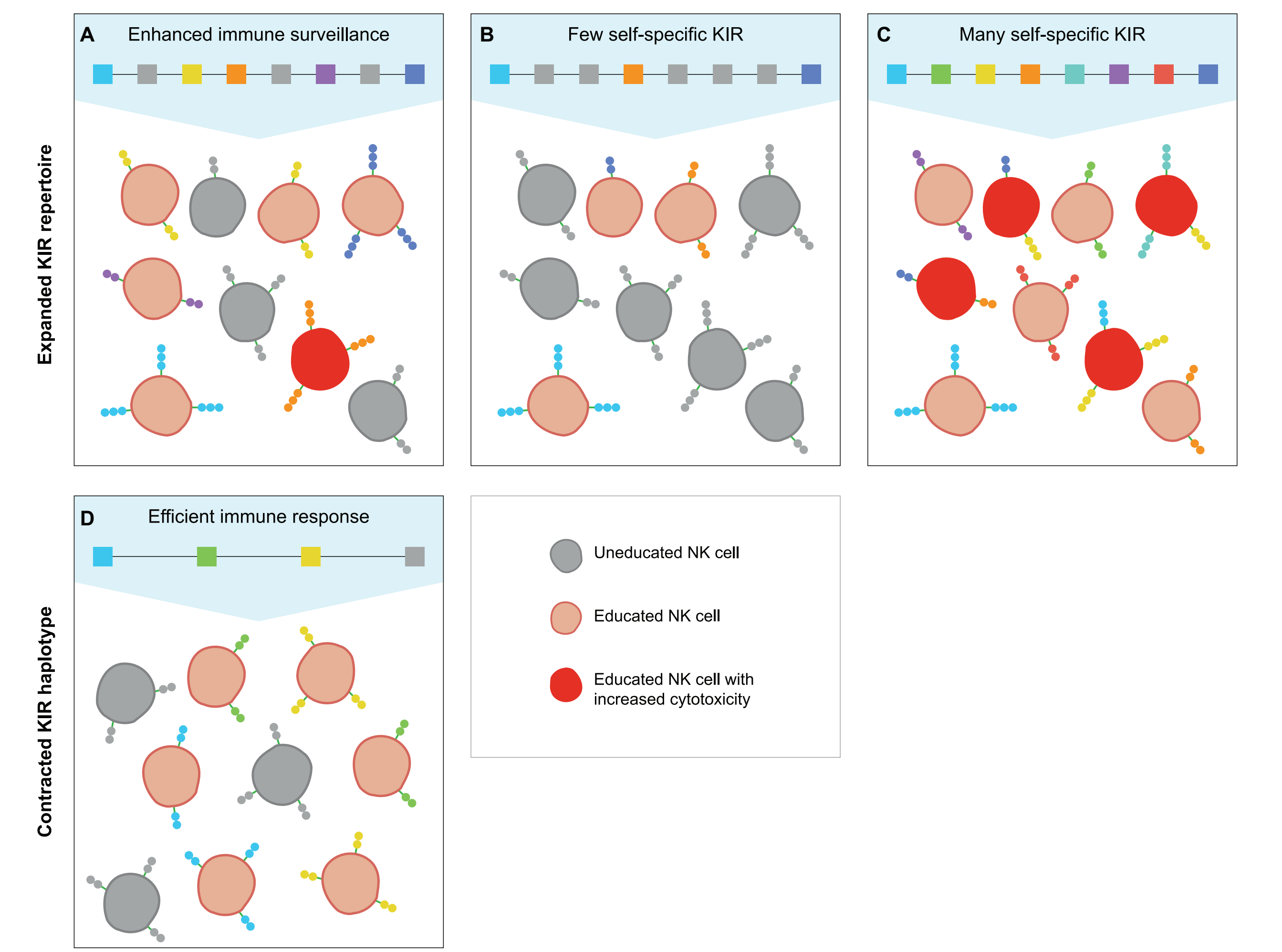


**FIGURE 4 |** The education of NK cells by expanded and contracted KIR haplotypes. The proposed populations of NK cells that are educated (red cells) by self-specific KIR (colored boxes and receptors) on expanded and contracted haplotypes. The gray boxes represent non-self-specific KIR, and are displayed on the uneducated NK cell clones (gray cells). **(A)** An expanded KIR haplotype provides a broad MHC class I specificity, and might educate some NK cell clones through multiple self-specific receptors, which increases their cytotoxicity (dark red cells). **(B)** A large KIR repertoire with only a few self-specific KIR might lower the fraction of educated NK cells and thereby provide an insufficient immune response, **(C)** whereas abundant self-specific receptors increase the educated fraction and their cytotoxicity. **(D)** A small KIR repertoire might educate large fractions of NK cells with a limited MHC class I specificity, which might provide largely sufficient immune responses.

balance might be reflected in the slightly variable *KIR* gene content per haplotype.

A large KIR repertoire is likely to provide a broad array of MHC class I specificities that may result in the education of an increased fraction of NK cells (**Figure 4A**). Moreover, the expression of multiple self-specific inhibitory KIR receptors by NK cell clones enhances the magnitude of their effector response (140). Although only a small population of NK cells dominantly expresses more than one inhibitory KIR receptor, an expanded KIR repertoire might enlarge this NK cell population size and elevate the strength of the NK cell response (**Figures 4A,C**). A potential detrimental effect of an expanded KIR haplotype might emerge if the repertoire comprises only a few or abundant self-specific receptors. On the one hand, the variegated expression of a large KIR repertoire that consists of few self-specific receptors might thin out the educated NK cell population and provide an inefficient immune surveillance (**Figure 4B**). Indications for a biased expression of self-specific KIR suggest modulation of the KIR expression by an individual's MHC class I repertoire (62, 140, 141), which would ensure a more robust immune response and might compensate for a large non-self-specific KIR expansion. On the other hand, a large repertoire of self-specific KIR might enlarge the fraction of educated NK cells that display increased activity, which might be protective in infections and cancer (**Figure 4C**). However, elevated NK cell activity, which might be further enhanced by the expression of multiple self-specific KIR on NK cell subsets, or excessive NK cell inhibition by abundant self-specific KIR interactions are also associated with implantation failure and recurrent miscarriages (142–145). Furthermore, overactivation might desensitize NK cells and result in hyporeactivity (146), which might weaken subsequent immune responses. Therefore, a large KIR repertoire that is used in NK cell education might act as a double-edged sword that can both enhance and compromise an individual's immune response.

In contrast, individuals that have a limited KIR haplotype rely on only one or few self-specific KIR receptors to educate their NK cells (**Figure 4D**). Even though a sufficient percentage of NK cells might be educated by a limited KIR repertoire, the specificity is restricted, and specialised NK cell populations might be lacking. The complete absence of NK cell education occurs in MHC class I-deficient mice, which display a near normal NK cell count with an overall reduced responsiveness (147, 148). In humans and other hominoid species, individuals that completely lack self-specific KIR are not documented. This indicates that even minimal KIR haplotypes provide education, and suggests that framework KIR receptors could play a substantial role in the NK cell education of hominoids. In addition, the chance that an individual completely lacks self-specific KIR receptors is reduced by the heterozygous nature of the *KIR* gene cluster. As far as we know, only few human and no non-human primate individuals are documented that were homozygous for their *KIR* haplotypes at an allele level (149). In a rhesus macaque family studied, one individual was assumed to be *KIR*-homozygous according to segregation. However, more detailed analysis illustrated that one *KIR* gene copy appeared to have gained point mutations that resulted in the haplotypes diverging at an allele level (29). This individual macaque possessed a largely homozygous KIR content, but did not display an impaired immune system; it also produced healthy offspring, which suggests that *KIR*-heterozygosity is not vital. However, *KIR* haplotype diversity might compensate for limited *KIR* haplotypes and improve the immune surveillance, as is also described for MHC heterozygosity (150–152).

In contrast to non-self-specific T lymphocytes, which are depleted upon a failed positive or negative selection in the thymus, uneducated NK cells are present in the peripheral blood. The relatively high level of uneducated NK cells in individuals with small or large non-self-specific KIR repertoires could affect their immune surveillance, but does not preclude an efficient immune response during infection or tumor formation. In fact, unlicensed NK cells appear to be more efficient at eradicating infected or malignant cells that persistently express MHC class I molecules or viral mimic ligands through their reactivation by cytokines or NKG2D receptors (153–155). Therefore, a fraction of uneducated NK cells in combination with a largely educated NK cell population might be more protective than a completely educated NK cell pool with broad *MHC class I* specificity.

There could be another factor, however, that limits expansion of the KIR haplotypes and gene repertoire, in addition to their role in NK cell education. In orangutans, MHC-B allotypes contain a -21M leader peptide, which would suggest education via the conserved CD94:NKG2A pathway (127). In contrast to macaques, the orangutan KIR system is not extensively expanded, and is more in line with other hominoids that display a KIR-dependent NK cell education. The emergence of MHC-C as a specialized ligand for KIR might override the dimorphism and coherent increase in MHC-E expression, and drive NK cell education via the KIR receptors. In addition, the number of characterized MHC-B molecules in orangutans is relatively low (IPD-MHC, release 3.4.0.1) (156). A larger sample group of orangutans or additional functional studies would be required to test our hypothesis for the differential KIR expansion in primate species that exert a KIR-independent or -dependent NK cell education.

Nevertheless, the diverse *KIR* haplotype content and overall gene repertoire in hominoids and Old World monkeys are likely to affect the education, activity, and function of their NK cells, but the precise effect of the haplotype expansions and contractions remains ambiguous. The equal distribution of both small and large KIR repertoires in humans and macaques indicates a balancing selection, which might be an ongoing process to achieve a haplotype equilibrium that serves differential functions, such as fighting infections and promoting successful pregnancy.

## CONCLUSION

The *KIR* gene system is well studied in humans, and reveals multiple mechanisms that contribute to the plasticity of this immunogenetic cluster (**Figure 2**). In other hominoid species, such as chimpanzees and orangutans, indications for a similar diversifying genetic toolset is evident, although robust data on some mechanisms are lacking, such as alternative splicing and variegated expression. The variability of the extensively diversified *KIR* gene cluster in macaques exceeds that observed in hominoids, with a prominent expansion of the lineage II *KIR* genes, which is largely mediated by recombination events. The rapid evolution of the *KIR* gene cluster may counteract the adaptive nature of pathogens. The species-specific diversification of the *KIR* gene cluster might be largely driven by co-evolution with their diversified MHC class I repertoire and thereby indirectly by the arms race with pathogens. In addition, a KIR-dependent or -independent NK cell education might impact the variable haplotype content and the extent of *KIR* gene expansion. Nevertheless, the different molecular mechanisms responsible for diversification of the *KIR* gene cluster are shared in Old World monkeys and hominoids, which suggests an evolutionary effort to diversify the *KIR* gene system.

## AUTHOR CONTRIBUTIONS

JB wrote the manuscript. NG and RB edited the manuscript. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

We thank D. Devine for editing the manuscript and F. van Hassel for preparing the figures.

## REFERENCES

1. Ljunggren HG, Karre K. In search of the 'missing self': MHC molecules and NK cell recognition. *Immunol Today*. (1990) 11:237–44. doi: 10.1016/0167-5699(90)90097-s
2. Vivier E, Tomasello E, Baratin M, Walzer T, Ugolini S. Functions of natural killer cells. *Nat Immunol*. (2008) 9:503–10.
3. Björkström NK, Riese P, Heuts F, Andersson S, Fauriat C, Ivarsson MA, et al. Expression patterns of NKG2A, KIR, and CD57 define a process of CD56dim NK-cell differentiation uncoupled from NK-cell education. *Blood*. (2010) 116:3853–64. doi: 10.1182/blood-2010-04-281675
4. Brodin P, Karre K, Hoglund P. NK cell education: not an on-off switch but a tunable rheostat. *Trends Immunol*. (2009) 30:143–9. doi: 10.1016/j.it.2009.01.006
5. Kelley J, Walter L, Trowsdale J. Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet*. (2005) 1:129–39. doi: 10.1371/journal.pgen.0010027
6. Martin AM, Freitas EM, Witt CS, Christiansen FT. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics*. (2000) 51:268–80. doi: 10.1007/s002510050620
7. Bruijnesteijn J, de Groot NG, Otting N, Maccari G, Guethlein LA, Robinson J, et al. Nomenclature report for killer-cell immunoglobulin-like receptors (KIR) in macaque species: new genes/alleles, renaming recombinant entities and IPD-NHKIR updates. *Immunogenetics*. (2020) 72:37–47. doi: 10.1007/s00251-019-01135-8
8. Marsh SG, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics*. (2003) 55:220–6.
9. Robinson J, Guethlein LA, Maccari G, Blokhuis J, Bimber BN, de Groot NG, et al. Nomenclature for the KIR of non-human species. *Immunogenetics*. (2018) 70:571–83. doi: 10.1007/s00251-018-1064-4
10. Sambrook JG, Bashirova A, Andersen H, Piatak M, Vernikos GS, Coggill P, et al. Identification of the ancestral killer immunoglobulin-like receptor gene in primates. *BMC Genomics*. (2006) 7:209. doi: 10.1186/1471-2164-7-209
11. Sanderson ND, Norman PJ, Guethlein LA, Ellis SA, Williams C, Breen M, et al. Definition of the cattle killer cell Ig-like receptor gene family: comparison with aurochs and human counterparts. *J Immunol*. (2014) 193:6016–30. doi: 10.4049/jimmunol.1401980
12. Guethlein LA, Abi-Rached L, Hammond JA, Parham P. The expanded cattle KIR genes are orthologous to the conserved single-copy KIR3DX1 gene of primates. *Immunogenetics*. (2007) 59:517–22. doi: 10.1007/s00251-007-0214-x
13. Wroblewski EE, Parham P, Guethlein LA. Two to tango: co-evolution of hominid natural killer cell receptors and MHC. *Front Immunol*. (2019) 10:177. doi: 10.3389/fimmu.2019.00177
14. de Groot NG, Heijmans CMC, van der Wiel MKH, Blokhuis JH, Mulder A, Guethlein LA, et al. Complex MHC class I gene transcription profiles and their functional impact in orangutans. *J Immunol*. (2016) 196:750. doi: 10.4049/jimmunol.1500820
15. Karl JA, Bohn PS, Wiseman RW, Nimityongskul FA, Lank SM, Starrett GJ, et al. Major histocompatibility complex class I haplotype diversity in Chinese rhesus macaques. *G3*. (2013) 3:1195–201. doi: 10.1534/g3.113.006254
16. Doxiadis GGM, de Groot N, Otting N, Blokhuis JH, Bontrop RE. Genomic plasticity of the MHC class I A region in rhesus macaques: extensive haplotype diversity at the population level as revealed by microsatellites. *Immunogenetics*. (2011) 63:73–83. doi: 10.1007/s00251-010-0486-4
17. Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE. Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res*. (2004) 14:1501–15. doi: 10.1101/gr.2134504
18. de Groot NG, Blokhuis JH, Otting N, Doxiadis GG, Bontrop RE. Co-evolution of the MHC class I and KIR gene families in rhesus macaques: ancestry and plasticity. *Immunol Rev*. (2015) 267:228–45. doi: 10.1111/imr.12313
19. Otting N, Heijmans CMC, Noort RC, de Groot NG, Doxiadis GGM, van Rood JJ, et al. Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc Natl Acad Sci USA*. (2005) 102:1626–31. doi: 10.1073/pnas.0409084102
20. Wiseman RW, Karl JA, Bohn PS, Nimityongskul FA, Starrett GJ, O'Connor DH. Haplessly hoping: macaque major histocompatibility complex made easy. *ILAR J*. (2013) 54:196–210. doi: 10.1093/ilar/ilt036
21. Schafer JL, Colantonio AD, Neidermyer WJ, Dudley DM, Connole M, O'Connor DH, et al. KIR3DL01 recognition of Bw4 ligands in the rhesus macaque: maintenance of Bw4 specificity since the divergence of apes and Old World monkeys. *J Immunol*. (2014) 192:1907–17. doi: 10.4049/jimmunol.1302883
22. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol*. (2013) 13:133–44. doi: 10.1038/nri3370
23. Rosner C, Kruse PH, Hermes M, Otto N, Walter L. Rhesus macaque inhibitory and activating KIR3D interact with Mamu-A-encoded ligands. *J Immunol*. (2011) 186:2156–63. doi: 10.4049/jimmunol.1002634
24. Bimber BN, Evans DT. The killer-cell immunoglobulin-like receptors of macaques. *Immunol Rev*. (2015) 267:246–58. doi: 10.1111/imr.12329
25. Hermes M, Weil S, Groth A, Dressel R, Koch J, Walter L. Characterisation of mouse monoclonal antibodies against rhesus macaque killer immunoglobulin-like receptors KIR3D. *Immunogenetics*. (2012) 64:845–8. doi: 10.1007/s00251-012-0640-2
26. Moreland AJ, Guethlein LA, Reeves RK, Broman KW, Johnson RP, Parham P, et al. Characterization of killer immunoglobulin-like receptor genetics and comprehensive genotyping by pyrosequencing in rhesus macaques. *BMC Genomics*. (2011) 12:295. doi: 10.1186/1471-2164-12-295
27. Wright PW, Li H, Huehn A, O'Connor GM, Cooley S, Miller JS, et al. Characterization of a weakly expressed KIR2DL1 variant reveals a novel upstream promoter that controls KIR expression. *Genes Immun*. (2014) 15:440–8. doi: 10.1038/gene.2014.34
28. Chan HW, Kurago ZB, Stewart CA, Wilson MJ, Martin MP, Mace BE, et al. DNA methylation maintains allele-specific KIR gene expression in human natural killer cells. *J Exp Med*. (2003) 197:245–55. doi: 10.1084/jem.20021127
29. Bruijnesteijn J, de Groot N, van der Wiel MKH, Otting N, de Vos-Rouweler AJM, de Groot NG, et al. Unparalleled rapid evolution of KIR genes in rhesus and cynomolgus macaque populations. *J Immunol*. (2020) 204:1770–86. doi: 10.4049/jimmunol.1901140
30. Martin AM, Kulski JK, Witt C, Pontarotti P, Christiansen FT. Leukocyte Ig-like receptor complex (LRC) in mice and men. *Trends Immunol*. (2002) 23:81–8. doi: 10.1016/s1471-4906(01)02155-x
31. Sambrook JG, Bashirova A, Palmer S, Sims S, Trowsdale J, Abi-Rached L, et al. Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Res*. (2005) 15:25–35. doi: 10.1101/gr.2381205
32. Wende H, Colonna M, Ziegler A, Volz A. Organization of the leukocyte receptor cluster (LRC) on human chromosome 19q13.4. *Mamm Genome*. (1999) 10:154–60. doi: 10.1007/s003359900961
33. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet*. (2013) 14:301–23. doi: 10.1146/annurev-genom-091212-153455
34. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet*. (2010) 19:737–51. doi: 10.1093/hmg/ddp538
35. Guethlein LA, Older Aguilar AM, Abi-Rached L, Parham P. Evolution of killer cell Ig-like receptor (KIR) genes: definition of an orangutan KIR haplotype reveals expansion of lineage III KIR associated with the emergence of MHC-C. *J Immunol*. (2007) 179:491–504. doi: 10.4049/jimmunol.179.1.491
36. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci USA*. (2000) 97:4778. doi: 10.1073/pnas.080588597
37. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet*. (2003) 73:823–34. doi: 10.1086/378594
38. Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet*. (2006) 79:41–53. doi: 10.1086/504600
39. Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, et al. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet*. (2007) 3:e184. doi: 10.1371/journal.pgen.0030184
40. Roe D, Vierra-Green C, Pyo CW, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun*. (2017) 18:127–34. doi: 10.1038/gene.2017.10

41. Rajalingam R, Gardiner CM, Canavez F, Vilches C, Parham P. Identification of seventeen novel KIR variants: fourteen of them from two non-Caucasian donors. *Tissue Antigens*. (2001) 57:22–31. doi: 10.1034/j.1399-0039.2001.057001022.x
42. Pyo C-W, Wang R, Vu Q, Cereb N, Yang SY, Duh F-M, et al. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics*. (2013) 14:89. doi: 10.1186/1471-2164-14-89
43. Bruijnesteijn J, van der Wiel MKH, Swelsen WTN, Otting N, de Vos-Rouweler AJM, Elferink D, et al. Human and rhesus macaque KIR haplotypes defined by their transcriptomes . *J Immunol*. (2018) 200:1692–701.
44. Vierra-Green C, Roe D, Hou L, Hurley CK, Rajalingam R, Reed E, et al. Allele-level haplotype frequencies and pairwise linkage disequilibrium for 14 KIR Loci in 506 European-American individuals. *PLoS One*. (2012) 7:e47491. doi: 10.1371/journal.pone.0047491
45. Norman PJ, Abi-Rached L, Gendzekhadze K, Hammond JA, Moesta AK, Sharma D, et al. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, and haplotypes. *Genome Res*. (2009) 19:757–69. doi: 10.1101/gr.085738.108
46. Leaton LA, Shortt J, Kichula KM, Tao S, Nemat-Gorgani N, Mentzer AJ, et al. Conservation, extensive heterozygosity, and convergence of signaling potential all indicate a critical role for KIR3DL3 in higher primates. *Front Immunol*. (2019) 10:24. doi: 10.3389/fimmu.2019.00024
47. Rajalingam R, Hong M, Adams EJ, Shum BP, Guethlein LA, Parham P. Short KIR haplotypes in pygmy chimpanzee (Bonobo) resemble the conserved framework of diverse human KIR haplotypes. *J Exp Med*. (2001) 193:135–46. doi: 10.1084/jem.193.1.135
48. de Groot NG, Heijmans CMC, Helsen P, Otting N, Pereboom Z, Stevens JMG, et al. Limited MHC class I intron 2 repertoire variation in bonobos. *Immunogenetics*. (2017) 69:677–88. doi: 10.1007/s00251-017-1010-x
49. Maibach V, Vigilant L. Reduced bonobo MHC class I diversity predicts a reduced viral peptide binding ability compared to chimpanzees. *BMC Evol Biol*. (2019) 19:14. doi: 10.1186/s12862-019-1352-0
50. Wroblewski EE, Guethlein LA, Norman PJ, Li Y, Shaw CM, Han AS, et al. Bonobos maintain immune system diversity with three functional types of MHC-B. *J Immunol*. (2017) 198:3480. doi: 10.4049/jimmunol.1601955
51. de Groot NG, Stevens JMG, Bontrop RE. Does the MHC confer protection against malaria in bonobos? *Trends Immunol*. (2018) 39:768–71. doi: 10.1016/j.it.2018.07.004
52. Gonzalez-Galarza FF, McCabe A, Santos E, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res*. (2020) 48:D783–8.
53. Nakimuli A, Chazara O, Farrell L, Hiby SE, Tukwasibwe S, Knee O, et al. Killer cell immunoglobulin-like receptor (KIR) genes and their HLA-C ligands in a Ugandan population. *Immunogenetics*. (2013) 65:765–75. doi: 10.1007/s00251-013-0724-7
54. Abi-Rached L, Moesta AK, Rajalingam R, Guethlein LA, Parham P. Human-specific evolution and adaptation led to major qualitative differences in the variable receptors of human and chimpanzee natural killer cells. *PLoS Genet*. (2010) 6:e1001192. doi: 10.1371/journal.pgen.1001192
55. Guethlein LA, Norman PJ, Heijmans CMC, de Groot NG, Hilton HG, Babrzadeh F, et al. Two orangutan species have evolved different KIR alleles and haplotypes. *J Immunol*. (2017) 198:3157–69. doi: 10.4049/jimmunol.1602163
56. NCBI *Macaca Mulatta Isolate AG07107 Chromosome 19, Whole Genome Shotgun Sequence*. (2018). Available online at: https://www.ncbi.nlm.nih.gov/nuccore/CM014354 (accessed July 01, 2020).
57. Blokhuis JH, van der Wiel MK, Doxiadis GGM, Bontrop RE. The mosaic of KIR haplotypes in rhesus macaques. *Immunogenetics*. (2010) 62:295–306. doi: 10.1007/s00251-010-0434-3
58. VandenBussche CJ, Dakshanamurthy S, Posch PE, Hurley CK. A single polymorphism disrupts the killer Ig-like receptor 2DL2/2DL3 D1 domain. *J Immunol*. (2006) 177:5347–57. doi: 10.4049/jimmunol.177.8.5347
59. Pando MJ, Gardiner CM, Gleimer M, McQueen KL, Parham P. The protein made from a common allele of KIR3DL1. (3DL1*004) is poorly expressed at cell surfaces due to substitution at positions 86 in Ig domain 0 and 182 in Ig domain 1. *J Immunol*. (2003) 171:6640–9. doi: 10.4049/jimmunol.171.12.6640
60. Frazier WR, Steiner N, Hou L, Dakshanamurthy S, Hurley CK. Allelic variation in KIR2DL3 generates a KIR2DL2-like receptor with increased binding to its HLA-C ligand. *J Immunol*. (2013) 190:6198–208. doi: 10.4049/jimmunol.1300464
61. Hilton HG, Norman PJ, Nemat-Gorgani N, Goyos A, Hollenbach JA, Henn BM, et al. Loss and gain of natural killer cell receptor function in an african hunter-gatherer population. *PLoS Genet*. (2015) 11:e1005439. doi: 10.1371/journal.pgen.1005439
62. Yawata M, Yawata N, Draghi M, Little AM, Partheniou F, Parham P. Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. *J Exp Med*. (2006) 203:633–45. doi: 10.1084/jem.20051884
63. Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, et al. Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nat Genet*. (2007) 39:733–40. doi: 10.1038/ng2035
64. Gardiner CM, Guethlein LA, Shilling HG, Pando M, Carr WH, Rajalingam R, et al. Different NK cell surface phenotypes defined by the DX9 antibody are due to KIR3DL1 gene polymorphism. *J Immunol*. (2001) 166:2992–3001. doi: 10.4049/jimmunol.166.5.2992
65. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SGE. IPD–the immuno polymorphism database. *Nucleic Acids Res*. (2013) 41:D1234–40.
66. Cooper DN. Functional intronic polymorphisms: buried treasure awaiting discovery within our genes. *Hum Genomics*. (2010) 4:284–8. doi: 10.1186/1479-7364-4-5-284
67. Toscano C, Klein K, Blievernicht J, Schaeffeler E, Saussele T, Raimundo S, et al. Impaired expression of CYP2D6 in intermediate metabolizers carrying the *41 allele caused by the intronic SNP 2988G> A: evidence for modulation of splicing events. *Pharmacogenet Genomics*. (2006) 16:755–66. doi: 10.1097/01.fpc.0000230112.96086.e0
68. Thomas R, Thio CL, Apps R, Qi Y, Gao X, Marti D, et al. A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection. *J Virol*. (2012) 86:6979–85. doi: 10.1128/jvi.00406-12
69. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*. (2011) 12:756–66. doi: 10.1038/nrg3098
70. Hwang DG, Green P. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*. (2004) 101:13994–4001. doi: 10.1073/pnas.0404142101
71. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics*. (2000) 156:297–304.
72. Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet*. (2001) 17:481–5. doi: 10.1016/s0168-9525(01)02409-x
73. Lercher MJ, Hurst LD. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*. (2002) 18:337–40. doi: 10.1016/s0168-9525(02)02669-0
74. Duret L, Arndt PF. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. (2008) 4:e1000071. doi: 10.1371/journal.pgen.1000071
75. Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*. (2008) 455:105–8. doi: 10.1038/nature07175
76. Zhu L, Wang Q, Tang P, Araki H, Tian D. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol Biol Evol*. (2009) 26:2353–61. doi: 10.1093/molbev/msp144
77. McDonald MJ, Wang W-C, Huang H-D, Leu J-Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol*. (2011) 9:e1000622. doi: 10.1371/journal.pbio.1000622
78. Maxwell LD, Williams F, Gilmore P, Meenagh A, Middleton D. Investigation of killer cell immunoglobulin-like receptor gene diversity: II. KIR2DS4. *Hum Immunol*. (2004) 65:613–21. doi: 10.1016/j.humimm.2004.02.028
79. Bruijnesteijn J, van der Wiel MKH, de Groot N, Otting N, de Vos-Rouweler AJM, Lardy NM, et al. Extensive alternative splicing of KIR transcripts. *Front Immunol*. (2018) 9:2846. doi: 10.3389/fimmu.2018.02846
80. Vilches C, Rajalingam R, Uhrberg M, Gardiner CM, Young NT, Parham P. KIR2DL5, a novel killer-cell receptor with a D0-D2 configuration of Ig-like domains. *J Immunol*. (2000) 164:5797. doi: 10.4049/jimmunol.164.11.5797

81. Dohring C, Samaridis J, Colonna M. Alternatively spliced forms of human killer inhibitory receptors. *Immunogenetics*. (1996) 44:227–30. doi: 10.1007/s002510050116
82. Wilson MJ, Torkar M, Trowsdale J. Genomic organization of a human killer cell inhibitory receptor gene. *Tissue Antigens*. (1997) 49:574–9. doi: 10.1111/j.1399-0039.1997.tb02804.x
83. Blokhuis JH, Doxiadis GG, Bontrop RE. A splice site mutation converts an inhibitory killer cell Ig-like receptor into an activating one. *Mol Immunol*. (2009) 46:640–8. doi: 10.1016/j.molimm.2008.08.270
84. Lee Y, Rio DC. Mechanisms and regulation of alternative Pre-mRNA splicing. *Annu Rev Biochem*. (2015) 84:291–323.
85. Goodridge JP, Lathbury LJ, Steiner NK, Shulse CN, Pullikotil P, Seidah NG, et al. Three common alleles of KIR2DL4 (CD158d) encode constitutively expressed, inducible and secreted receptors in NK cells. *Eur J Immunol*. (2007) 37:199–211. doi: 10.1002/eji.200636316
86. Rajalingam R, Parham P, Abi-Rached L. Domain shuffling has been the main mechanism forming new hominoid killer cell Ig-like receptors. *J Immunol*. (2004) 172:356–69. doi: 10.4049/jimmunol.172.1.356
87. Valiante NM, Uhrberg M, Shilling HG, Lienert-Weidenbach K, Arnett KL, D'Andrea A, et al. Functionally and structurally distinct NK cell receptor repertoires in the peripheral blood of two human donors. *Immunity*. (1997) 7:739–51. doi: 10.1016/s1074-7613(00)80393-3
88. Andersson S, Malmberg JA, Malmberg KJ. Tolerant and diverse natural killer cell repertoires in the absence of selection. *Expe Cell Res*. (2010) 316:1309–15. doi: 10.1016/j.yexcr.2010.02.030
89. Santourlidis S, Trompeter HI, Weinhold S, Eisermann B, Meyer KL, Wernet P, et al. Crucial role of DNA methylation in determination of clonally distributed killer cell Ig-like receptor expression patterns in NK cells. *J Immunol*. (2002) 169:4253–61. doi: 10.4049/jimmunol.169.8.4253
90. Stewart CA, Van Bergen J, Trowsdale J. Different and divergent regulation of the KIR2DL4 and KIR3DL1 promoters. *J Immunol*. (2003) 170:6073–81. doi: 10.4049/jimmunol.170.12.6073
91. Trompeter HI, Gomez-Lozano N, Santourlidis S, Eisermann B, Wernet P, Vilches C, et al. Three structurally and functionally divergent kinds of promoters regulate expression of clonally distributed killer cell Ig-like receptors (KIR), of KIR2DL4, and of KIR3DL3. *J Immunol*. (2005) 174:4135–43. doi: 10.4049/jimmunol.174.7.4135
92. Li H, Pascal V, Martin MP, Carrington M, Anderson SK. Genetic control of variegated KIR gene expression: polymorphisms of the Bi-directional KIR3DL1 promoter are associated with distinct frequencies of gene expression. *PLoS Genet*. (2008) 4:e1000254. doi: 10.1371/journal.pgen.1000254
93. Davies GE, Locke SM, Wright PW, Li H, Hanson RJ, Miller JS, et al. Identification of bidirectional promoters in the human KIR genes. *Genes Immun*. (2007) 8:245–53. doi: 10.1038/sj.gene.6364381
94. Li H, Wright PW, McCullen M, Anderson SK. Characterization of KIR intermediate promoters reveals four promoter types associated with distinct expression patterns of KIR subtypes. *Genes Immun*. (2016) 17:66–74. doi: 10.1038/gene.2015.56
95. Parham P. Immunogenetics of killer-cell immunoglobulin-like receptors. *Tissue Antigens*. (2003) 62:194–200. doi: 10.1034/j.1399-0039.2003.00126.x
96. Gomez-Lozano N, Trompeter HI, de Pablo R, Estefania E, Uhrberg M, Vilches C. Epigenetic silencing of potentially functional KIR2DL5 alleles: implications for the acquisition of KIR repertoires by NK cells. *Eur J Immunol*. (2007) 37:1954–65. doi: 10.1002/eji.200737277
97. Cisneros E, Moraru M, Gómez-Lozano N, López-Botet M, Vilches C. KIR2DL5: an orphan inhibitory receptor displaying complex patterns of polymorphism and expression. *Front Immunol*. (2012) 3:289. doi: 10.3389/fimmu.2012.00289
98. Vilches C, Gardiner CM, Parham P. Gene structure and promoter variation of expressed and nonexpressed variants of the *KIR2DL5*. Gene. *J Immunol*. (2000) 165:6416. doi: 10.4049/jimmunol.165.11.6416
99. Gomez-Lozano N, Estefania E, Williams F, Halfpenny I, Middleton D, Solis R, et al. The silent KIR3DP1 gene (CD158c) is transcribed and might encode a secreted receptor in a minority of humans, in whom the KIR3DP1, KIR2DL4 and KIR3DL1/KIR3DS1 genes are duplicated. *Eur J Immunol*. (2005) 35:16–24. doi: 10.1002/eji.200425493
100. van Bergen J, Stewart CA, van den Elsen PJ, Trowsdale J. Structural and functional differences between the promoters of independently expressed killer cell Ig-like receptors. *Eur J Immunol*. (2005) 35:2191–9. doi: 10.1002/eji.200526201
101. Marquardt N, Scharenberg M, Mold JE, Hård J, Kekäläinen E, Buggert M, et al. High-dimensional analysis reveals a distinct population of adaptive-like tissue-resident NK cells in human lung. *bioRxiv* [Preprint]. (2019). doi: 10.1101/2019.12.20.883785
102. Marquardt N, Beziat V, Nystrom S, Hengst J, Ivarsson MA, Kekalainen E, et al. Cutting edge: identification and characterization of human intrahepatic CD49a+ NK cells. *J Immunol*. (2015) 194:2467–71. doi: 10.4049/jimmunol.1402756
103. Ivarsson MA, Stiglund N, Marquardt N, Westgren M, Gidlof S, Bjorkstrom NK. Composition and dynamics of the uterine NK cell KIR repertoire in menstrual blood. *Mucosal Immunol*. (2017) 10:322–31. doi: 10.1038/mi.2016.50
104. Björkström NK, Béziat V, Cichocki F, Liu LL, Levine J, Larsson S, et al. CD8 T cells express randomly selected KIRs with distinct specificities compared with NK cells. *Blood*. (2012) 120:3455–65. doi: 10.1182/blood-2012-03-416867
105. Mingari MC, Moretta A, Moretta L. Regulation of KIR expression in human T cells: a safety mechanism that may impair protective T-cell responses. *Immunol Today*. (1998) 19:153–7. doi: 10.1016/s0167-5699(97)01236-x
106. Huard B, Karlsson L. KIR expression on self-reactive CD8+ T cells is controlled by T-cell receptor engagement. *Nature*. (2000) 403:325–8. doi: 10.1038/35002105
107. Liu Y, Kuick R, Hanash S, Richardson B. DNA methylation inhibition increases T cell KIR expression through effects on both promoter methylation and transcription factors. *Clin Immunol*. (2009) 130:213–24. doi: 10.1016/j.clim.2008.08.009
108. Lam TH, Shen M, Chia JM, Chan SH, Ren EC. Population-specific recombination sites within the human MHC region. *Heredity*. (2013) 111:131–8. doi: 10.1038/hdy.2013.27
109. Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, et al. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am J Hum Genet*. (2005) 76:634–46. doi: 10.1086/429393
110. Voorter CE, Gerritsen KE, Groeneweg M, Wieten L, Tilanus MG. The role of gene polymorphism in HLA class I splicing. *Int J Immunogenet*. (2016) 43:65–78. doi: 10.1111/iji.12256
111. Krangel MS. Secretion of HLA-A and -B antigens via an alternative RNA splicing pathway. *J Exp Med*. (1986) 163:1173–90. doi: 10.1084/jem.163.5.1173
112. Dunn PP, Hammond L, Coates E, Street J, Griner L, Darke C. A "silent" nucleotide substitution in exon 4 is responsible for the "alternative expression" of HLA-A*01:01:38L through aberrant splicing. *Hum Immunol*. (2011) 72:717–22.
113. Dai Z-X, Zhang G-H, Zhang X-H, Xia H-J, Li S-Y, Zheng Y-T. The β2-microglobulin–free heterodimerization of rhesus monkey MHC Class I a with its normally spliced variant reduces the ubiquitin-dependent degradation of MHC class I A. *J Immunol*. (2012) 2012:1100665.
114. Dai ZX, Zhang GH, Zhang XH, Zheng YT. Identification and characterization of a novel splice variant of rhesus macaque MHC IA. *Mol Immunol*. (2013) 53:206–13. doi: 10.1016/j.molimm.2012.08.006
115. Ishitani A, Geraghty DE. Alternative splicing of HLA-G transcripts yields proteins with primary structures resembling both class I and class II antigens. *Proc Natl Acad Sci USA*. (1992) 89:3947. doi: 10.1073/pnas.89.9.3947
116. Rouas-Freiss N, Bruel S, Menier C, Marcou C, Moreau P, Carosella ED. Switch of HLA-G alternative splicing in a melanoma cell line causes loss of HLA-G1 expression and sensitivity to NK lysis. *Int J Cancer*. (2005) 117:114–22. doi: 10.1002/ijc.21151
117. Kirszenbaum M, Moreau P, Gluckman E, Dausset J, Carosella E. An alternatively spliced form of HLA-G mRNA in human trophoblasts and evidence for the presence of HLA-G transcript in adult lymphocytes. *Proc Natl Acad Sci USA*. (1994) 91:4209. doi: 10.1073/pnas.91.10.4209
118. Rizzo R, Bortolotti D, Bolzani S, Fainardi E. HLA-G molecules in autoimmune diseases and infections. *Front Immunol*. (2014) 5:592. doi: 10.3389/fimmu.2014.00592

119. Fleming KA, McMichael A, Morton JA, Woods J, McGee JO. Distribution of HLA class 1 antigens in normal human tissue and in mammary cancer. *J Clin Pathol*. (1981) 34:779–84. doi: 10.1136/jcp.34.7.779

120. Wei X, Orr HT. HLA Class I. 2nd ed. In: Delves PJ editor. *Encyclopedia of Immunology*. Oxford: Elsevier (1998). p. 1108–11.

121. Dellgren C, Nehlin JO, Barington T. Cell surface expression level variation between two common human leukocyte antigen alleles, HLA-A2 and HLA-B8, is dependent on the structure of the C terminal part of the alpha 2 and the alpha 3 domains. *PLoS One*. (2015) 10:e0135385. doi: 10.1371/journal.pone.0135385

122. Apps R, Meng Z, Del Prete GQ, Lifson JD, Zhou M, Carrington M. Relative expression levels of the HLA class-I proteins in normal and HIV-infected cells. *J Immunol*. (2015) 194:3594–600. doi: 10.4049/jimmunol.1403234

123. Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, et al. Influence of HLA-C expression level on HIV control. *Science*. (2013) 340:87.

124. Otting N, Heijmans CM, van der Wiel M, de Groot NG, Doxiadis GG, Bontrop RE. A snapshot of the Mamu-B genes and their allelic repertoire in rhesus macaques of Chinese origin. *Immunogenetics*. (2008) 60:507–14. doi: 10.1007/s00251-008-0311-5

125. He Y, Tian Z. NK cell education via nonclassical MHC and non-MHC ligands. *Cell Mol Immunol*. (2017) 14:321–30. doi: 10.1038/cmi.2016.26

126. He Y, Peng H, Sun R, Wei H, Ljunggren H-G, Yokoyama WM, et al. Contribution of inhibitory receptor TIGIT to NK cell education. *J Autoimmun*. (2017) 81:1–12. doi: 10.1016/j.jaut.2017.04.001

127. Horowitz A, Djaoud Z, Nemat-Gorgani N, Blokhuis J, Hilton HG, Béziat V, et al. Class I HLA haplotypes form two schools that educate NK cells in different ways. *Sci Immunol*. (2016) 1:eaag1672. doi: 10.1126/sciimmunol.aag1672

128. Anfossi N, André P, Guia S, Falk CS, Roetynck S, Stewart CA, et al. Human NK cell education by inhibitory receptors for MHC Class I. *Immunity*. (2006) 25:331–42.

129. Borrego F, Ulbrecht M, Weiss EH, Coligan JE, Brooks AG. Recognition of human histocompatibility leukocyte antigen (HLA)-E complexed with HLA class I signal sequence-derived peptides by CD94/NKG2 confers protection from natural killer cell-mediated lysis. *J Exp Med*. (1998) 187:813–8. doi: 10.1084/jem.187.5.813

130. Braud VM, Allan DS, O'Callaghan CA, Soderstrom K, D'Andrea A, Ogg GS, et al. HLA-E binds to natural killer cell receptors CD94/NKG2A. B and C. *Nature*. (1998) 391:795–9. doi: 10.1038/35869

131. Lee N, Llano M, Carretero M, Ishitani A, Navarro F, López-Botet M, et al. HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. *Proc Natl Acad Sci USA*. (1998) 95:5199–204. doi: 10.1073/pnas.95.9.5199

132. Fauriat C, Ivarsson MA, Ljunggren HG, Malmberg KJ, Michaëlsson J. Education of human natural killer cells by activating killer cell immunoglobulin-like receptors. *Blood*. (2010) 115:1166–74. doi: 10.1182/blood-2009-09-245746

133. Lee N, Goodlett DR, Ishitani A, Marquardt H, Geraghty DE. HLA-E surface expression depends on binding of TAP-dependent peptides derived from certain HLA class I signal sequences. *J Immunol*. (1998) 160:4951–60.

134. Malnati MS, Peruzzi M, Parker KC, Biddison WE, Ciccone E, Moretta A, et al. Peptide specificity in the recognition of MHC class I by natural killer cell clones. *Science*. (1995) 267:1016–8. doi: 10.1126/science.7863326

135. Rajagopalan S, Long EO. The direct binding of a p58 killer cell inhibitory receptor to human histocompatibility leukocyte antigen (HLA)-Cw4 exhibits peptide selectivity. *J Exp Med*. (1997) 185:1523–8. doi: 10.1084/jem.185.8.1523

136. Hansasuta P, Dong T, Thananchai H, Weekes M, Willberg C, Aldemir H, et al. Recognition of HLA-A3 and HLA-A11 by KIR3DL2 is peptide-specific. *Eur J Immunol*. (2004) 34:1673–9. doi: 10.1002/eji.200425089

137. Fadda L, Borhis G, Ahmed P, Cheent K, Pageon SV, Cazaly A, et al. Peptide antagonism as a mechanism for NK cell activation. *Proc Natl Acad Sci USA*. (2010) 107:10160. doi: 10.1073/pnas.0913745107

138. Li Y, Mariuzza RA. Structural basis for recognition of cellular and viral ligands by NK cell receptors. *Front Immunol*. (2014) 5:123. doi: 10.3389/fimmu.2014.00123

139. Carrillo-Bustamante P, de Boer RJ, Keşmir C. Specificity of inhibitory KIRs enables NK cells to detect changes in an altered peptide environment. *Immunogenetics*. (2018) 70:87–97. doi: 10.1007/s00251-017-1019-1

140. Yu J, Heller G, Chewning J, Kim S, Yokoyama WM, Hsu KC. Hierarchy of the human natural killer cell response is determined by class and quantity of inhibitory receptors for self-HLA-B and HLA-C ligands. *J Immunol*. (2007) 179:5977–89. doi: 10.4049/jimmunol.179.9.5977

141. Shilling HG, Young N, Guethlein LA, Cheng NW, Gardiner CM, Tyan D, et al. Genetic control of human NK cell repertoire. *J Immunol*. (2002) 169:239–47. doi: 10.4049/jimmunol.169.1.239

142. Templer S, Sacks G. A blessing and a curse: is high NK cell activity good for health and bad for reproduction? *Hum Fertil*. (2016) 19:166–72. doi: 10.1080/14647273.2016.1219072

143. Shakhar K, Rosenne E, Loewenthal R, Shakhar G, Carp H, Ben-Eliyahu S. High NK cell activity in recurrent miscarriage: what are we really measuring? *Hum Reprod*. (2006) 21:2421–5. doi: 10.1093/humrep/del131

144. Colucci F. The role of KIR and HLA interactions in pregnancy complications. *Immunogenetics*. (2017) 69:557–65. doi: 10.1007/s00251-017-1003-9

145. Hiby SE, Regan L, Lo W, Farrell L, Carrington M, Moffett A. Association of maternal killer-cell immunoglobulin-like receptors and parental HLA-C genotypes with recurrent miscarriage. *Hum Reprod*. (2008) 23:972–6. doi: 10.1093/humrep/den011

146. Tripathy SK, Keyel PA, Yang L, Pingel JT, Cheng TP, Schneeberger A, et al. Continuous engagement of a self-specific activation receptor induces NK cell tolerance. *J Exp Med*. (2008) 205:1829–41. doi: 10.1084/jem.20072446

147. Liao NS, Bix M, Zijlstra M, Jaenisch R, Raulet D. MHC class I deficiency: susceptibility to natural killer (NK) cells and impaired NK activity. *Science*. (1991) 253:199–202. doi: 10.1126/science.1853205

148. Hoglund P, Ohlen C, Carbone E, Franksson L, Ljunggren HG, Latour A, et al. Recognition of beta 2-microglobulin-negative (beta 2m-) T-cell blasts by natural killer cells from normal but not from beta 2m- mice: nonresponsiveness controlled by beta 2m- bone marrow in chimeric mice. *Proc Natl Acad Sci USA*. (1991) 88:10332–6. doi: 10.1073/pnas.88.22.10332

149. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA Class I genotypes at highest resolution via high-throughput sequencing. *Am J Hum Genet*. (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023

150. Doherty PC, Zinkernagel RM. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*. (1975) 256:50–2. doi: 10.1038/256050a0

151. McClelland EE, Penn DJ, Potts WK. Major histocompatibility complex heterozygote superiority during coinfection. *Infect Immun*. (2003) 71:2079–86. doi: 10.1128/iai.71.4.2079-2086.2003

152. Penn DJ, Damjanovich K, Potts WK. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci USA*. (2002) 99:11260–4. doi: 10.1073/pnas.162006499

153. Tu MM, Mahmoud AB, Makrigiannis AP. Licensed and unlicensed NK cells: differential roles in cancer and viral control. *Front Immunol*. (2016) 7:166. doi: 10.3389/fimmu.2016.00166

154. Tarek N, Le Luduec JB, Gallagher MM, Zheng J, Venstrom JM, Chamberlain E, et al. Unlicensed NK cells target neuroblastoma following anti-GD2 antibody treatment. *J Clin Investig*. (2012) 122:3260–70. doi: 10.1172/jci62749

155. Orr MT, Murphy WJ, Lanier LL. 'Unlicensed' natural killer cells dominate the response to cytomegalovirus infection. *Nature Immunol*. (2010) 11:321–7. doi: 10.1038/ni.1849

156. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J, et al. IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. *Nucleic Acids Res*. (2017) 45:D860–4.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Immunology

ORIGINAL RESEARCH
published: 23 September 2020
doi: 10.3389/fimmu.2020.02136

# A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus

**Oscar L. Rodriguez[1], William S. Gibson[2], Tom Parks[3], Matthew Emery[1], James Powell[1], Maya Strahl[1], Gintaras Deikus[1], Kathryn Auckland[3], Evan E. Eichler[4,5], Wayne A. Marasco[6], Robert Sebra[1,7], Andrew J. Sharp[1], Melissa L. Smith[1,2,7]*, Ali Bashir[1]* and Corey T. Watson[2]***

*[1] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [2] Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, United States, [3] Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [4] Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, United States, [5] Howard Hughes Medical Institute, University of Washington, Seattle, WA, United States, [6] Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Department of Medicine, Harvard Medical School, Boston, MA, United States, [7] Icahn Institute of Data Science and Genomic Technology, New York, NY, United States*

An incomplete ascertainment of genetic variation within the highly polymorphic immunoglobulin heavy chain locus (IGH) has hindered our ability to define genetic factors that influence antibody-mediated processes. Due to locus complexity, standard high-throughput approaches have failed to accurately and comprehensively capture IGH polymorphism. As a result, the locus has only been fully characterized two times, severely limiting our knowledge of human IGH diversity. Here, we combine targeted long-read sequencing with a novel bioinformatics tool, IGenotyper, to fully characterize IGH variation in a haplotype-specific manner. We apply this approach to eight human samples, including a haploid cell line and two mother-father-child trios, and demonstrate the ability to generate high-quality assemblies (>98% complete and >99% accurate), genotypes, and gene annotations, identifying 2 novel structural variants and 15 novel IGH alleles. We show multiplexing allows for scaling of the approach without impacting data quality, and that our genotype call sets are more accurate than short-read (>35% increase in true positives and >97% decrease in false-positives) and array/imputation-based datasets. This framework establishes a desperately needed foundation for leveraging IG genomic data to study population-level variation in antibody-mediated immunity, critical for bettering our understanding of disease risk, and responses to vaccines and therapeutics.

**Keywords: immunoglobulin heavy chain locus, single nucleotide variation, structural variation, antibody, B cell receptor, long-read sequencing**

OPEN ACCESS

**Edited by:**
*Martin Maiers,
National Marrow Donor Program,
United States*

**Reviewed by:**
*Gur Yaari,
Bar-Ilan University, Israel
Guido Moll,
Charité – Universitätsmedizin Berlin,
Germany*

***Correspondence:**
*Melissa L. Smith
mlsmit53@louisville.edu
Ali Bashir
ali.bashir@gmail.com
Corey T. Watson
corey.watson@louisville.edu*

**Specialty section:**
*This article was submitted to
Alloimmunity and Transplantation,
a section of the journal
Frontiers in Immunology*

**Received:** *10 June 2020*
**Accepted:** *06 August 2020*
**Published:** *23 September 2020*

**Citation:**
*Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, Deikus G, Auckland K, Eichler EE, Marasco WA, Sebra R, Sharp AJ, Smith ML, Bashir A and Watson CT (2020) A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. Front. Immunol. 11:2136. doi: 10.3389/fimmu.2020.02136*

## INTRODUCTION

Defining the factors that contribute to differences in the antibody (Ab) response is critical to furthering our understanding of immunological diseases, and informing the design of vaccines and therapeutics. Antibodies are an extremely diverse protein family in humans, encoded by >100 highly homologous gene segments that reside within one of the most structurally complex and polymorphic regions of the human genome (1–3). The immunoglobulin heavy chain (IGH) locus, specifically, consists of >50 variable (IGHV), >20 diversity (IGHD), 6 joining (IGHJ), and 9 constant (IGHC) functional/open reading frame (F/ORF) genes that encode the heavy chains of expressed Abs (4). Greater than 250 IGH gene segment alleles are curated in the ImMunoGeneTics Information System (IMGT) database (4), and this number continues to increase (2, 5–11). The locus is enriched for single nucleotide variants (SNVs) and large structural variants (SVs) involving functional genes (5, 12–19), at which allele frequencies are known to vary among human populations (2, 19, 20). The formation of the Ab repertoire is mediated by several complex molecular processes and can be influenced by many factors. Studies in twins have demonstrated that features of the Ab repertoire are heritable, and IG germline variants have been shown to directly impact antibody usage and antigen-specificity (20–22). Together this highlights the need to better understand the genetic factors that contribute to variation in antibody-mediated immunity.

Currently, existing genomic resources and tools for the IG loci are incomplete and poorly represent germline diversity across human populations. Historically, the complexity of the IGH locus has hindered our ability to comprehensively characterize polymorphisms within this region using high-throughput approaches (3, 23). In fact, only two complete haplotypes in IGH have been fully resolved and characterized (1, 2). As a result, IGH has been largely overlooked by genome-wide studies, leaving our understanding of the contribution of polymorphisms within IGH in antibody-mediated immunity incomplete (2, 3, 23). While early studies uncovered associations to disease susceptibility within IGH, few links have been made by genome-wide association studies (GWAS) and whole genome sequencing (WGS) (3, 24, 25). Moreover, little is known about the genetic regulation of the human Ab response.

To define the role of IGH variation in Ab function and disease, all classes of variation must be resolved (5, 13, 16, 20, 26, 27). Although approaches have been developed for utilizing genomic or Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) data, variant calling and broad-scale haplotype inference are restricted primarily to coding regions (7, 8, 17–19, 28). To fully characterize genetic diversity in the IG loci, specialized genotyping methods capable of capturing locus-wide polymorphism with nucleotide resolution are required. Indeed, such methods have been applied to better resolve complex and hyper-polymorphic immune loci elsewhere in the genome (29, 30).

Long-read sequencing technologies have been used to detect chromosomal rearrangements (31), novel SVs (32, 33), and SVs missed by standard short-read sequencing methods (34, 35), including applications in the complex killer immunoglobulin-like receptors (KIR) (36) and human leukocyte antigen (HLA) (31, 37, 38) loci. Furthermore, the sensitivity of SV detection is improved by resolving variants in a haplotype-specific manner (35, 39). When long-read sequencing has been combined with specific target enrichment methods, using either a CRISPR/Cas9 system (40, 41) or DNA probes (42, 43), it has been shown to yield accurate and contiguous assemblies. Targeted approaches have enabled higher resolution genotyping of the HLA loci (44) and KIR regions (45, 46).

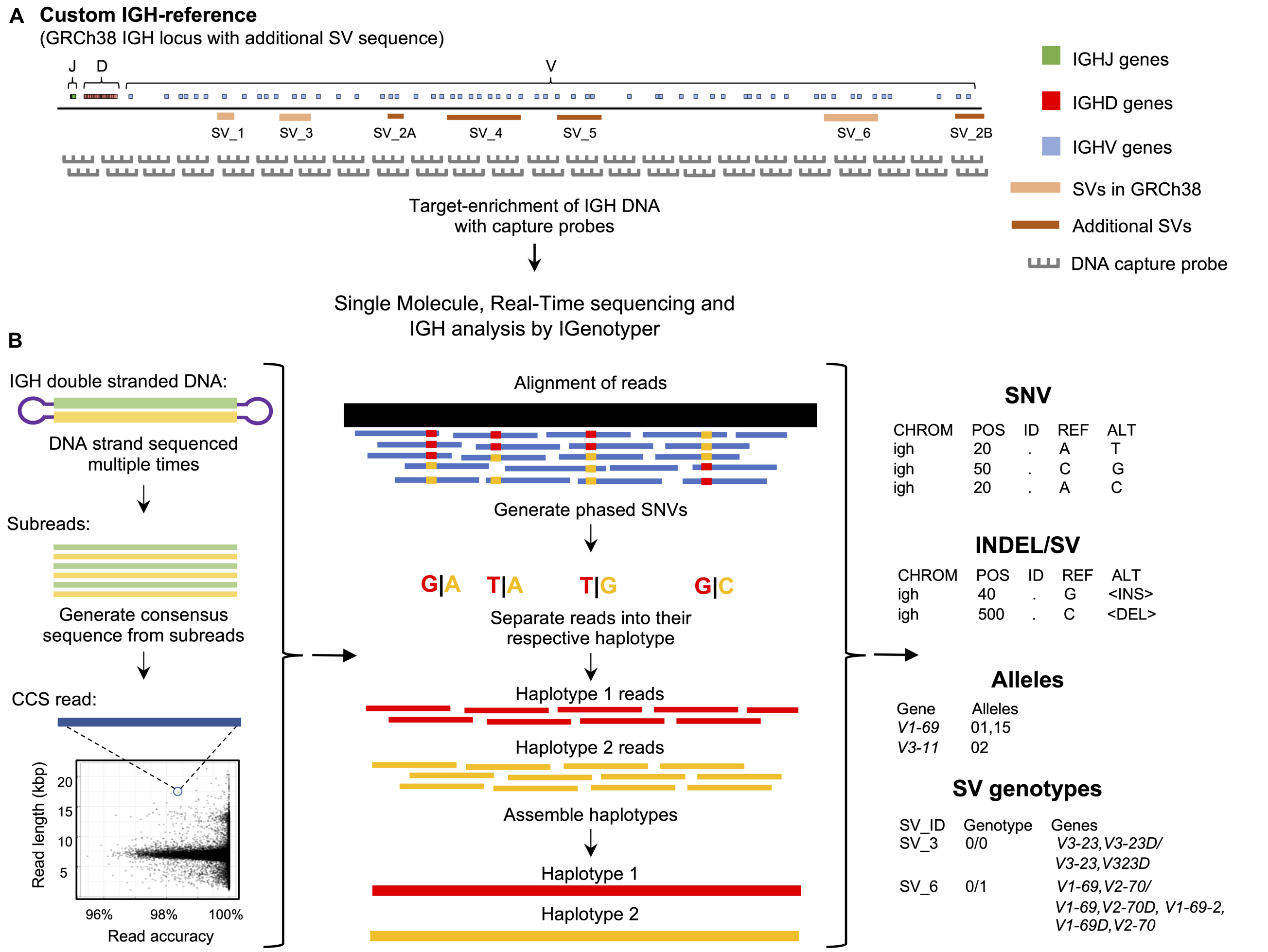
Here, we present a novel framework that utilizes IGH-targeted long-read sequencing, paired with a new IG genomics analysis tool, IGenotyper[1], to characterize variation in the IGH locus (**Figure 1**). We apply this strategy to eight human samples, and leverage orthogonal data and pedigree information for benchmarking and validation. We demonstrate that our approach leads to hiqh-quality assemblies across the IGH locus, allowing for comprehensive genotyping of SNVs, insertions and deletions (indels), SVs, as well as annotation of IG gene segments, alleles, and associated non-coding elements. We show that genotype call sets from our pipeline are more comprehensive than those generated using alternative short-read and array-/imputation-based methods. Finally, we demonstrate that use of long-range phasing/haplotype information improves assembly contiguity, and that sample multiplexing can be employed to scale the approach in a cost-effective manner without impacting data quality.

## MATERIALS AND METHODS

### Library Preparation and Sequencing

Genomic DNA samples were procured from Coriell Repositories (1000 Genomes Project, 1KGP donors; Camden, NJ, United States) and collaborators. Briefly, 1–2 micrograms of high molecular weight genomic DNA from each sample was sheared using g-tube to ∼8 Kb (Covaris, Woburn, MA, United States). These sheared gDNA samples were size selected to include 5–9 Kb fragments using a BluePippin (Sage Science, Beverly, MA, United States). Following size selection, each sample was End Repaired and A-tailed following the standard KAPA library preparation protocol (Roche, Basel, Switzerland). For multiplexed samples, adapters containing sequence barcodes (Pacific Biosciences, Menlo Park, CA, United States) and a universal priming sequence were ligated onto each sample. Each sample was PCR amplified for 9 cycles using HS LA Taq (Takara, Mountain View, CA, United States) and cleaned with 0.7X AMPure beads to remove small fragments and excess reagents (Beckman Coulter, Brea, CA, United States). The genomic DNA libraries were then captured following the SeqCap protocol which was modified to increase final capture reaction volume by 1.5X (Roche, Basel, Switzerland). Following capture, the libraries were washed following the SeqCap protocol, substituting vortexing with gentle flicking. The washed capture libraries were

[1] https://igenotyper.github.io/



SNV

| CHROM | POS | ID | REF | ALT |
|---|---|---|---|---|
| igh | 20 | . | A | T |
| igh | 50 | . | C | G |
| igh | 20 | . | A | C |

INDEL/SV

| CHROM | POS | ID | REF | ALT |
|---|---|---|---|---|
| igh | 40 | . | G | <INS> |
| igh | 500 | . | C | <DEL> |

Alleles

| Gene | Alleles |
|---|---|
| *V1-69* | 01,15 |
| *V3-11* | 02 |

SV genotypes

| SV_ID | Genotype | Genes |
|---|---|---|
| SV_3 | 0/0 | *V3-23, V3-23D/ V3-23, V323D* |
| SV_6 | 0/1 | *V1-69, V2-70/ V1-69, V2-70D, V1-69-2, V1-69D, V2-70* |

**FIGURE 1 |** An overview of the custom capture and IGenotyper workflow used to detect IGH variation in a haplotype-specific manner. **(A)** A schematic of the custom IGH-reference used in this study, which includes the GRCh38 IGH locus with the addition of known SV sequences inserted into their respective positions in the locus. Brown bars indicate the positions of inserted SVs in the IGH-reference; pink bars indicate the positions of additional known SVs present in GRCh38 relative to GRCh37. Positions of IGHJ (green), IGHD (red) and IGHV (blue) genes in the IGH-reference are also indicated. The oligo capture probes used in the panels designed for this study were based on the sequence of this custom IGH-reference. **(B)** After targeted long-read capture, constructed libraries undergo SMRT sequencing, and the resulting data is processed with IGenotyper. Raw SMRT sequences with at least 2 subreads are converted into higher accuracy CCS reads. Phased SNVs are identified from the CCS reads and used to phase both the CCS reads and subreads, which are then used to assemble haplotype-specific assemblies of IGH. SNVs, indels, SVs and IG gene/alleles are identified from the assembly and CCS reads. SVs embedded into the IGH-reference are genotyped using the assembly, CCS reads, and SNV calls.

PCR amplified for 18 cycles using HS LA Taq and cleaned with 0.7X AMPure beads.

Capture libraries were prepared for PacBio sequencing using the SMRTbell Template Preparation Kit 1.0 (Pacific Biosciences, Menlo Park, CA, United States). Briefly, each sample was treated with a DNA Damage Repair and End Repair mix in order to repair nicked DNA. SMRTbell adapters were ligated onto each capture library to complete SMRTbell construction. The SMRTbell libraries were then treated with exonuclease III and VII to remove any unligated gDNA and cleaned with 0.45X AMPure PB beads (Pacific Biosciences, Menlo Park, CA, United States). Resulting libraries were prepared for sequencing according to the manufacturer's protocol and sequenced as single libraries per SMRTcell with P6/C4 chemistry and 6 h movies on the RSII system, or as multiplexed libraries per SMRTcell 1M, annealed to primer V4 and sequenced using 3.0 chemistry and 20 h movies, on the Sequel system (see **Supplementary Table S4** for details).

## Creating a Custom IGH Locus Reference

The IGH locus, excluding the IGH constant gene region (chr14:106,326,710-107,349,540) was removed from the human genome reference build GRCh37, and the expanded custom IGH locus reference was inserted in its place. The expanded custom IGH locus includes sequence spanning IGH from the GRCh38 reference assembly, plus the addition of known SVs. The included SVs were previously characterized from fosmid clones AC244473.3, AC241995, AC234225, AC233755, KC162926, KC162924, AC231260, AC244456 and

KC162925, and sequence from human genome reference build GRCh37 (chr14:106,527,905-106,568,151). The IGenotyper toolkit command 'IG-make-ref' takes as input the human genome reference build GRCh37 and creates the custom IGH locus reference.

## IGenotyper: A Streamlined Analysis Tool for IGH Locus Assembly, Variant Detection/Genotyping, and Gene Feature Annotation

Running IGenotyper returns multiple output files: (1) the alignment of the circular consensus sequence (CCS) reads and assembled locus to the reference in BAM format; (2) the assembled IGH locus in FASTA format, (3) the SNVs in VCF; (4) indels and SVs in BED format; (5) a parsable file with genotyped SVs; (6) a parsable file with the detected alleles for each functional/ORF gene; and (7) several tab delimited files detailing different sequencing run and assembly statistics. The BAM file contains phased CCS reads and includes haplotype annotation in the read group tag of every read. This allows the user to separate the reads into their respective haplotype in the Integrative Genomics Viewer (IGV) visualization tool (47). The VCF file contains annotations indicating whether SNVs reside within SV regions, and IG gene features, including coding, intron, leader part 1 (LP1), and recombination signal (RS) sequences. A user-friendly summary file is produced with links to output files (see **Supplementary Figure S1**; sample summary output for NA19240), including summary tables and figures pertaining to: locus sequence coverage; counts of SNVs, indels and SVs; allele annotations/genotypes for each IGHV, IGHD, and IGHJ genes; and lists of novel alleles.

### Running IGenotyper

IGenotyper has three main commands: 'phase', 'assemble', and 'detect'. The input is the subread bam output from the RSII or Sequel sequencing run. 'phase' phases the subreads and CCS reads. 'assemble' partitions the IGH locus into haplotype-specific regions and assembles each region. 'detect' detects SNVs, indels and SVs, genotypes 5 SVs embedded in the IGH reference and assigns the IGH genes to alleles from the IGMT database.

### Phasing SMRT Sequencing Reads

Raw SMRT sequences with at least 2 subreads are converted into consensus sequences using the 'ccs' command[2]. CCS reads and subreads are aligned to our expanded custom IGH locus reference using BLASR (48). Phased SNVs are detected from the CCS reads using the WhatsHap (49) 'find_snv_candidates', 'genotype', and 'phase' commands. The subreads and CCS reads are phased using the command 'phase-bam' from MsPAC (50).

### Assembling the IGH Locus

Haplotype blocks are defined using the WhatsHap 'stats' command. Haplotype-specific CCS reads within each haplotype block are assembled separately using Canu (51); regions outside haplotype blocks are assembled using all aligned CCS reads. Regions lacking contigs recruit raw subreads and repeat the assembly process. Contigs with a quality score less than 20 are filtered.

### Detecting Variants From SNVs to SVs

SNVs are detected from the assembly aligned to the reference. SNVs in the VCF are annotated with:

1. Contig id used to detect the SNV.
2. Overlapping SV id (if any).
3. A true or false value if the SNV is also detected by the CCS reads.
4. Whether the SNV falls within an intron, leader part 1 sequence or gene exon.
5. Whether the SNV is within the IGHV, IGHD, or IGHJ region.
6. For phased blocks, the haplotype block id and genotype.

Indels and SVs are detected using the 'sv-calling' command from MsPAC (50). Importantly, each indel and SV is sequence-resolved since they are identified from a multiple sequence alignment using the haplotype-specific assemblies and reference.

### Assigning Alleles to IGH Genes Extracted From the Assembly

Assembled sequences overlapping the IGH genes are extracted and compared to the alleles in the IMGT database (v202031-2). Sequences not observed in the database are labeled as novel. CCS reads overlapping IGH gene sequences are also extracted and compared to the IMGT database. To provide supporting evidence for the allele, a count for the number of CCS reads with the same sequence found in the assembly is reported (currently only supported for Sequel data). Assembly and CCS sequences are outputted in a FASTA file; gene names, alleles (and allele sequence if novel), and read support are outputted in a tab-delimited file.

## Validating IGenotyper Assemblies

### Assessing the Accuracy of the NA12878 and NA19240 Assemblies

The accuracy of assemblies for NA19240 and NA12878 was assessed by BLAST (v2.7.1+) alignment of contigs to fosmids assembled from Sanger sequencing (NA19240 accession numbers: AC241513.1, AC234301.2, KC162926.1, AC233755.2, AC234135.3, AC244463.2; NA12878 accession numbers: AC245090.1, AC244490.2).

The accuracy of the assemblies was also assessed using Illumina data. Illumina HiSeq 2500 2 × 126 paired-end, PCR-free sequencing data (ERR894723, ERR894724, ERR899709, ERR899710 and ERR899711) was downloaded from the European Nucleotide Archive and aligned to the NA19240 assembly using bwa mem (v0.7.15-r1140). The total coverage across all sequencing runs was 75.6x and Pilon (v1.23) reported an accuracy of 99.996 (102 errors in 2,396,307 bp). Additionally, Illumina NovaSeq 6000 2 × 151 paired-end, TruSeq PCR-free sequencing run ERR3239334 was downloaded from the European Nucleotide Archive and aligned to the NA12878 assembly. The coverage was 35x and Pilon reported an accuracy of 99.991 (167 errors in 1,918,794 bp).

[2] https://github.com/PacificBiosciences/ccs

## Manual Curation of IGHV3-30 and IGHV1-69 Gene Regions

*IGHV3-30* and *IGHV1-69* gene duplication regions did not completely assemble into a single contig per haplotype, but instead were split into multiple contigs. To resolve these regions an additional curation step was employed: contigs were aligned to each other using BLAST and overlapping contigs with high alignment score were merged.

In NA19240, the *IGHV3-30* gene region duplication was initially assembled into 8 contigs. Two contigs were merged to form a novel SV containing a ∼25 Kb deletion relative to the IGH-reference. The two contigs overlapped by 7,706 bp with 5 bp mismatches and 9 gaps (11 gap bases). The alternate haplotype was initially assembled into 6 contigs. The 6 contigs overlapped by more than 2.3 Kbp with 0 bp mismatches and a total of 8 gap bases, allowing them to be merged into a single contig. Both haplotypes were validated with fosmids and assemblies from the parents. The resulting contigs resolved the SVs on both haplotypes. This process was repeated for NA12878, and in both probands for the *IGHV1-69* gene region.

Leveraging parental and fosmid assembly data, we determined that NA19240 carried three distinct haplotypes within the SV region spanning *IGHV1-69*, *IGHV2-70D*, *IGHV1-69-2*, *IGHV1-69D*, and *IGHV2-70*. An insertion haplotype carrying all genes within the region was paternally inherited; a deletion haplotype, lacking *IGHV2-70D*, *IGHV1-69-2*, and *IGHV1-69D*, was inherited from the mother; and a second deletion haplotype was detected in both the capture/IGenotyper and fosmid assembly data, but was not supported by either parental dataset. This deletion haplotype was identical to the paternally derived insertion haplotype on the flanks of the deletion event, suggesting it represented a somatic SV. Whether this arose natively in NA19240 or is an artifact found only within the LCL is not known. To construct the most accurate assemblies of the inherited haplotypes, we attempted to remove reads representing this somatic deletion and performed a local reassembly. This allowed us to produce more accurate contigs across this region which exhibited higher concordance to both fosmid and parental datasets.

## Comparing Variants From IGenotyper to Other Datasets

### Comparing Illumina and IGenotyper SNV Calls in CHM1

SNVs from our assembly and Illumina reference alignment to GRCh37 were compared to a ground truth SNV dataset generated by aligning GRCh38 region chr14:106,329,408-107,288,965 to GRCh37 using pbmm2 (v1.0.0). Positions with base differences in the alignment were labeled as SNVs. 2 × 126 TruSeq PCR-free Illumina libraries (SRR3099549 and SRR2842672) were aligned to GRCh37 with bwa (0.7.15-r1140) and SNVs were detected using the standard protocol with GATK (v3.6) tools, HaplotypeCaller and GenotypeGVCFs (52). The SNVs were filtered using bcftools (v1.9) for SNVs with genotype quality greater than 60 and read depth greater than 10.

### Comparing 1000 Genomes Project SNVs to NA12878 and NA19240 SNVs Detected by IGenotyper

SNVs detected by IGenotyper were compared to SNVs from the 1KGP phase 3 dataset. IGH-specific SNVs from the 1KGP (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr14.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz,ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz, and ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA19240/cg_data/NA19240_lcl_SRR832874.wgs.COMPLETE_GENOMICS.20130401.snps_indels_svs_meis.high_coverage.genotypes.vcf.gz) were extracted using ‘bcftools view –output-type v, –regions 14:106405609-107349540, –min-ac 1, –types snps‘. Overlap between the 1KGP Phase 3 SNVs and SNVs detected by IGenotyper was determined using BEDTools ‘intersect‘ command. Overlapping SNVs with discordant genotypes were labeled as discordant/non-overlapping SNVs.

### Comparing Indels, SVs, and BioNano Data From the 1000 Genome Structural Variation Consortium

Indels and SVs detected by IGenotyper were compared to indels and SVs from the Human Genome Structural Variation (HGSV) Consortium (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/integration/20170515_Integrated_indels_Illumina_PacBio/Illumina_Indels_Merged_20170515.vcf.gz, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20180627_PhasedSVMSPAC/PhasedSVMsPAC.NA19240.vcf). Additionally, BioNano SV calls (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20180502_bionano/GM19240_DLE1_SV_hg38_indel.vcf and ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20170109_BioNano_SV_update/GM19240_YRI_Daughter_20170109_bionano_SVMerged_InDel.vcf.gz) were used to validate SVs identified by IGenotyper in NA19240.

### Analysis of SNVs in Regions Accessible to Next Generation Sequencing Methods

SNVs were evaluated to determine if they were within regions accessible by next generation sequencing methods. The “strict” accessibility track was converted to bed format from:

https://hgdownload.soe.ucsc.edu/gbdb/hg19/1000Genomes/phase3/20141020.strict_mask.whole_genome.bb.

The number of SNVs within the “strict” accessible regions was determined by using the BEDtools “intersect” command.

### Calculating Hardy-Weinberg Equilibrium at NA19240 and NA12878 SNVs

SNVs from the 1KGP (phase 3) were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr14.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz. Samples corresponding to the “EUR” superpopulation were extracted from the VCF file, and samples corresponding to the "AFR" superpopulation were extracted into a separate VCF file. Hardy-Weinberg Equilibrium (HWE) was calculated using vcftools with the option "—hardy."

# RESULTS

## Novel Tools for Comprehensively Characterizing IG Haplotype Diversity

To interrogate locus-wide IGH variants, we implemented a framework that pairs targeted DNA capture with single molecule, real time (SMRT) sequencing (Pacific Biosciences) (**Figure 1A**). Roche Nimblegen SeqCap EZ target-enrichment panels (Wilmington, MA, United States) were designed using DNA target sequences from the human IGH locus. Critically, rather than using only a single representative IGH haplotype (e.g., those available as part of either the GRCh37 or GRCh38 human reference assembly) we based our design on non-redundant sequences from the GRCh38 haplotype (2), as well as additional complex SV and insertion haplotypes distinct from GRCh38 (1, 2) (**Figure 1A** and **Supplementary Table S1**). Additional details, including the exact target sequences used and additional specifications of these capture panels are provided in the **Supplementary Material** (**Supplementary Note 1** and **Tables S2, S3**).

To process and analyze these long-read IGH genomic sequencing data, we developed IGenotyper (**Figure 1B**)[3] which utilizes and builds on existing tools to generate diploid assemblies across the IGHV, IGHD, and IGHJ regions (see section "Materials and Methods"); for ease, we refer to these three regions (excluding IGHC) collectively as IGH. From generated assemblies, IGenotyper additionally produces comprehensive summary reports of SNV, indel, and SV genotype call sets, as well as IG gene/allele annotations. For read mapping, SNV/indel/SV calling, and sequence annotation, the pipeline leverages a custom IGH locus reference that represents known SV sequences in a contiguous, non-redundant fashion (**Figure 1A**); this reference harbors the same sequence targets used for the design of target-enrichment panels, and ensures that known IGH SVs in the human population can be interrogated.

## Benchmarking Performance Using a Haploid DNA Sample

We first benchmarked our performance using genomic DNA from a complete haploid hydatidiform mole sample (CHM1), from which IGH had been previously assembled using Bacterial Artificial Chromosome (BAC) clones and Sanger sequencing (2). This reference serves as the representation of IGH in GRCh38. Using panel designs mentioned above, we prepared two SMRTbell libraries with an average insert size of 6–7.5 Kb for sequencing on both the RSII and Sequel systems (**Supplementary Table S4**). We observed a mean subread coverage across our custom IGH reference (**Figure 1A**) of 557.9x (RSII) and 12,006.4x (Sequel 1), and mean circular consensus sequence (CCS) read coverage of 45.1x (RSII) and 778.2x (Sequel 1). The average Sequel CCS phred quality score was 70 (99.999991% accuracy), with an average read length of 6,457 bp (**Supplementary Figure S2**). Noted differences in target-enrichment panels tested are described in **Supplementary Note 1**.

To most effectively use these data to assess IGenotyper performance, we combined reads from both libraries for assembly (**Supplementary Table S4**). A total of 970,302 bp (94.8%) of IGH (chr14:105,859,947-106,883,171; GRCh38) was spanned by >1,000× subread coverage, and 1,006,287 bp (98.3%) was spanned by >20× CCS coverage (**Figure 2A**). The mean CCS coverage spanning IGHV, IGHD, and IGHJ coding sequences was 160.3× (median = 42.5×; **Figure 2B**). Compared to GRCh38, IGenotyper assembled 1,009,792 bases (98.7%) of the IGH locus in the CHM1 dataset (**Figure 2C** and **Table 1**). Gap sizes between contigs ranged from 177 to 3,787 bp (median = 456 bp) in length. Only 37 (<0.004% of bases) single nucleotide differences were observed when compared to GRCh38 (base pair concordance >99.99%). In addition, 220 potential indel errors were identified (**Supplementary Table S5**). The majority of these (199/220) were 1–2 bp in length, 61.8% of which (123/199) occurred in homopolymer sequences, consistent with known sources of sequencing error in SMRT sequencing and other technologies (**Supplementary Table S6**). We also observed a 2,226 bp indel consisting of a 59mer tandem repeat motif near the gene *IGHV1-69* (**Supplementary Figure S3**). This tandem repeat was unresolved in GRCh38 (**Supplementary Figure S4** and Note 2), which was reconstructed using a Sanger shot-gun assembly approach (2); it remains unclear whether this event represents an improvement in the IGenotyper assembly over GRCh38, or is a sequencing/assembly artifact. Nonetheless, the total number of discordant bases associated with indels (2,521 bp) accounts for only <0.28% of the assembly.

All known SVs previously described in CHM1 (2) were present in the IGenotyper assembly, accounting for all IGHV ($n$ = 47), IGHD ($n$ = 27), and IGHJ ($n$ = 6) F/ORF gene segments in this sample. In addition to genes previously characterized by BAC sequencing, the IGenotyper assembly additionally spanned *IGHV7-81*. Alleles identified by IGenotyper were 100% concordant with those identified previously in GRCh38 (**Figure 2D**) (2).

## Assessing the Accuracy of Diploid IGH Assemblies

We next assessed the ability of IGenotyper to resolve diploid assemblies in IGH, using a Yoruban (YRI; NA19240, NA19238, NA1239) and European (CEU; NA12878, NA12891, NA12892) trio from the 1000 Genomes Project [1KGP (53); **Supplementary Figure S4** and **Supplementary Table S4**]. Lymphoblastoid cell lines (LCLs), which are the primary source of 1KGP sample DNA are known to harbor V(D)J somatic rearrangements within the IG loci, including reduced coverage in IGHD, IGHJ, and proximal IGHV regions (2, 19). However, because IGenotyper assembles the IGH locus in a local haplotype-specific manner, V(D)J rearrangements can be easily detected (**Supplementary Figures S5,S6** and **Table S7**). Nonetheless, we focused our

[3] http://igenotyper.github.io/

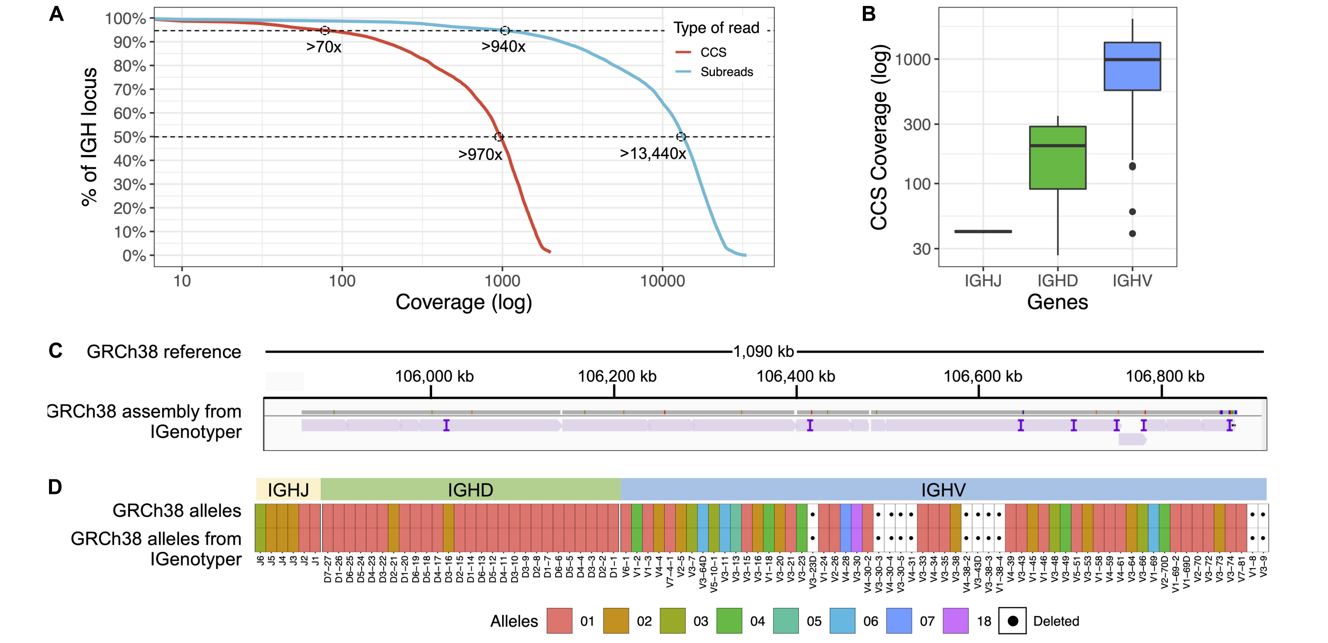


**FIGURE 2 |** Benchmarking targeted long-read sequencing and assembly in a haploid DNA sample. **(A)** The empirical cumulative subread (blue) and CCS (red) coverage in IGH from the combined CHM1 dataset. The subread coverage for 95 and 50% (dotted line) of the locus is greater than 940× and 13,440×, and the CCS read coverage for 95 and 50% of the locus is greater than 70× and 970×, respectively. **(B)** CCS coverage across IGHJ, IGHD and IGHV genes. The average CCS coverage of IGHV genes was >1,000×. **(C)** IGenotyper assembly of CHM1 aligned to GRCh38. Purple tick marks represent indels in the IGenotyper assembly relative to GRCh38. **(D)** IGHJ, IGHD, and IGHV alleles detected by IGenotyper in CHM1 compared to alleles previously annotated in GRCh38.

**TABLE 1 |** Assembly statistics and evaluation of the accuracy of the haplotype-specific assemblies.

| Sample | Contigs (n) | Assembly size (bp) | Assembly validation | | |
|---|---|---|---|---|---|
| | | | Concordance with fosmids (SMRT sequencing) | Concordance with BACs or fosmids (Sanger sequencing) | Concordance with Pilon/Illumina |
| CHM1 | 16 | 1,026,385 | NA | 99.996% | NA |
| NA19240 | 38 | 1,829,616 | 99.996% | 99.99% | 99.99% |
| NA12878 | 45 | 1,442,310 | 99.995% | 100.0% | 99.99% |

analysis exclusively on the IGHV region (9 Kb downstream of *IGHV6-1* to telomere) to avoid potential technical artifacts that would hinder our benchmarking assessment.

IGH enrichment was performed and libraries were sequenced on the RSII or Sequel platform (**Supplementary Table S4**). For diploid samples, IGenotyper (**Figure 1B**) first identifies haplotype blocks using CCS reads that span multiple heterozygous SNVs within a sample. Within each haplotype block, CCS reads are then partitioned into their respective haplotype and assembled independently to derive assembly contigs representing each haplotype in that individual. Reads within blocks of homozygosity that cannot be phased are collectively assembled, as these blocks are considered to represent either: (1) homozygous regions, in which both haplotypes in the individual are presumed to be identical, or (2) hemizygous regions, in which the individual is presumed to harbor an insertion or deletion on only one chromosome (**Supplementary Figure S7**).

We assessed IGenotyper performance in the probands of each trio. IGenotyper assemblies were composed of 41 and 49 haplotype blocks in NA19240 and NA12878, respectively (**Supplementary Table S8**). Of these, 20/41 and 24/49 blocks in each respective sample were identified as heterozygous, in which haplotype-specific assemblies could be generated, totaling 826,548 bp (69.28%) in NA19240, and 424,834 bp (35.61%) in NA12878. Within these heterozygous blocks, the mean number of heterozygous positions was 76.16 (NA19240) and 52.08 (NA12878). Summing the bases assembled across both heterozygous and homozygous/hemizygous contigs, complete assemblies comprised 1.8 Mb of diploid resolved sequence in NA19240 and 1.4 Mb in NA12878 (**Table 1**). The difference in size is partially due to V(DJ) rearrangements and large deletions present in NA12878 relative to NA19240 (**Figure 3**).

We next validated the accuracy of these assemblies using several orthogonal datasets: Sanger- and SMRT-sequenced

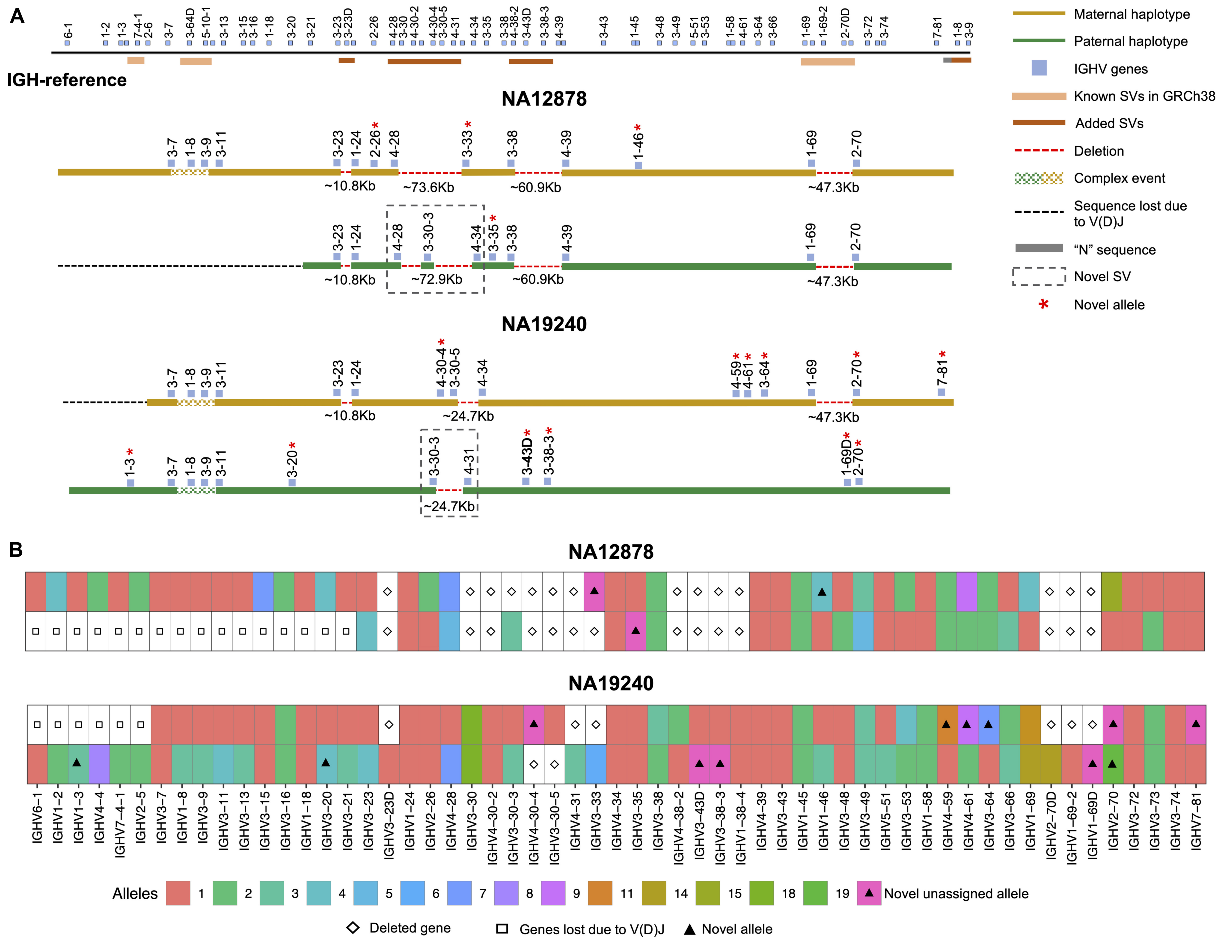


**FIGURE 3 |** Haplotype-resolved assembly for characterizing structural variants and IGHV gene alleles in NA12878 and NA19240. **(A)** A schematic of the custom IGH-reference spanning the IGHV gene region (top). Brown bars indicate the positions of inserted SVs in the IGH-reference; pink bars indicate the positions of additional known SVs present in GRCh38 relative to GRCh37. Positions of IGHV (blue) genes are also indicated. Schematic depictions of resolved maternally (gold) and paternally (green) inherited haplotypes in NA12878 and NA19240 are shown. Detected deletions within annotated SV regions are labeled with a red dotted line, with detected sizes of each event also provided. Genes directly flanking or within detected SVs are labeled. Genes with novel alleles are labeled with red asterisks, and novel SVs are indicated with dotted boxes. **(B)** Alleles predicted by IGenotyper for each gene across both haplotypes in NA12878 and NA19240. Novel alleles are marked by a filled triangle, and deleted genes and genes not present due to V(D)J recombination are marked by a diamond and square, respectively.

fosmid clones, and paired-end Illumina data (**Table 1**). The Sanger-sequenced fosmids (2) ($n = 6$, NA19240; $n = 2$, NA12878) spanned 210.4 Kb of the NA19240 IGenotyper assembly and 70.2 Kb of the NA12878 assembly (**Supplementary Table S9**). The percent identity relative to the Sanger-sequenced fosmids was 99.989% for NA19240 and 100% NA12878. We also compared IGenotyper assemblies to additional fosmid assemblies in these samples (Rodriguez et al., unpublished data) sequenced using SMRT sequencing ($n = 85$, NA19240; $n = 73$, NA12878). These collectively spanned 1.5 Mb (82%; NA19240) and 1.2 Mb (82%; NA12878) of the IGenotyper assemblies, aligning with 99.996 and 99.995% sequence identity, respectively (**Table 1**). The numbers of putative indel errors were 276 (NA19240) and 188 (NA12878) (**Supplementary Table S5**). Of these indels, 254/276 and 180/188 were 1–2 bp indels, 71.26% (181/254) and 63.33% (114/180) of which were within homopolymers (**Supplementary Table S6**). Finally, we additionally assessed assembly accuracy using publicly available 30× PCR-free paired-end Illumina TruSeq data from NA19240 and NA12878 (34, 53). We observed a total of 25 discordant bases and 143 indels in the NA19240 assembly (accuracy = 99.989%), and 45 discordant bases and 154 indels in the NA12878 assembly (accuracy = 99.991%; **Table 1**).

## Assessing Local Phasing Accuracy and Extending Haplotype-Specific Assemblies With Long-Range Phasing Information

To assess local phasing accuracy, we also profiled the IGH locus in the parents of NA19240 and NA12878 (**Supplementary Table S4**). IGenotyper uses read-backed phasing to delineate reads within haplotype blocks prior to assembly. We tested the accuracy of local phasing (variant phasing within each haplotype block) by comparing read-backed and trio-based phased genotypes. No phase-switch errors were observed in the heterozygous haplotype blocks ($n$ = 20 blocks, NA19240; $n$ = 24, NA12878). Within homozygous blocks (excluding known SV sites), 27/57,313 (0.05%; NA19240) and 23/139,029 (0.02%; NA12878) bases did not follow a Mendelian inheritance pattern (**Supplementary Table S10**).

In both NA19240 and NA12878, we observe low localized read coverage in various regions of the locus, representing known technical limitations of probe-based DNA capture (54). Because of this, as well as regions of homozygosity/hemizygosity, IGenotyper is limited in its ability to generate phased haplotype assemblies across the entirety of the locus. However, we reasoned that with long-range phase information (e.g., trio genotypes) contigs from an IGenotyper assembly can be assigned to parental haplotypes. To assess this, heterozygous SNVs in NA19240 and NA12878 were phased using both sequencing reads and parental SNVs, resulting in completely phased haplotypes. To determine potential impacts on accuracy using either the local or long-range phasing approach, we compared each assembly type in the probands. Only 12 (NA19240) and 7 (NA12878) base differences were found between the locally phased and long-range phased assemblies. Taken together, these data suggest that individual contig assemblies generated by IGenotyper have high phasing accuracy.

We anticipate that alternative forms of long-range phasing data will be available in the future. One example would be IGHV, IGHD, and IGHJ haplotype information inferred from AIRR-seq data (17, 18). We assessed whether AIRR-seq based haplotype inference could be theoretically applied, by identifying the number haplotype blocks harboring heterozygous IGHV gene segments. In NA19240 and NA12878, 10/20 and 6/24 of the assembled heterozygous contig blocks harbored at least one heterozygous IGHV gene. In total, this equated to 53.5% (442,057 bps) in NA19240 and 80.72% (342,942 bps) in NA12878 of heterozygous bases that could theoretically be linked using this type of coarse long-range phase information, highlighting the potential strength of pairing the two approaches in larger numbers of samples.

## IGenotyper Produces Accurate Gene Annotation, SNV, Indel, and SV Variant Call Sets From Diploid Assemblies

Previous studies have demonstrated that assembling diploid genomes in a haplotype-specific manner increases the accuracy of variant detection (34, 35, 39, 50, 55–57) and facilitates greater resolution on the full spectrum of variant classes (58). In addition to IGH locus assembly, IGenotyper detects SNVs, short indels, and SVs, including SNV calls within previously characterized complex SV/insertion regions. IGenotyper also provides direct genotypes for five previously described biallelic SVs (see **Supplementary Table S11**). This excludes the structurally complex *IGHV3-30* gene region, known to harbor multiple complex haplotypes; however, IGenotyper assemblies can be used for manual curation of this region (see below).

Using the fully phased diploid assemblies from each proband (**Figure 3**), we assessed the validity of annotations/variant calls. We compared proband gene annotations and variant call sets to fosmid and parental assembly data (**Table 2**). In each sample, we noted the presence of a V(D)J recombination event on one chromosome, which resulted in the artificial loss of alleles (**Figure 3A**). However, because these events were detectable, they did not preclude our ability to make accurate annotations and variant calls.

In NA19240, IGenotyper identified 79 unique non-redundant alleles across 57 IGHV genes (**Figure 3B** and **Supplementary Tables S12 and 13**); 12 of these alleles were not found in IMGT, representing novel alleles. All 79 alleles were validated by parental and/or fosmid assembly data (**Supplementary Table S12**). In NA12878, 56 non-redundant alleles were called at 44 IGHV genes (**Figure 3B** and **Supplementary Tables S14, 15**), three of which were novel; all 56 alleles were validated (**Supplementary Table S14**).

Across IGHV we identified 2,912 SNVs, 49 indels (2–49 bps), and 11 SVs (>50 bps) in NA19240. Collectively, IGenotyper-based genotypes for the parents of NA19240 and/or genotypes from the fosmids supported 2,869/2,912 SNVs, 31/36 indels, and 11/11 SVs in NA19240. In NA12878, we identified 2,329 SNVs, 36 indels (2–49 bps), and 3 SVs (>50 bps), 2,308 (SNVs), 20 (indels), and 3 (SVs) of which were supported by orthogonal data. Included in the SVs called from both probands were events within previously identified SV regions (**Figure 1A** and **Supplementary Table S11**). All of these regions are polymorphic at the population level (2, 20, 28), and several involve complex duplications and repeat structures (**Figure 3A**). Strong concordance was observed in these regions between proband IGenotyper assemblies, fosmid clones, and parental CCS reads/assemblies (**Supplementary Figures S8–S13** and Table S16).

Additionally, we discovered novel SVs in both NA12878 and NA19240 within the region spanning the genes *IGHV4-28* to *IGHV4-34* (**Figure 3A**). This site is a known hotspot of structural polymorphisms, in which six SV haplotypes have been fully or partially resolved (2). The longest haplotype characterized to date (2) contains four ∼25 Kb segmental duplication blocks. The novel SV haplotype in NA12878 contains a single ∼25 Kb segmental duplication block, and lacks 6 of

**TABLE 2** | Count of different variants identified by IGenotyper.

| Sample | Variant type | Count | Validation rate |
|---|---|---|---|
| NA19240 | SNV | 2,912 | 98.5% (2, 869/2, 912) |
| | Indel* | 49 | 100% (49/49) |
| | SV | 11 | 100% (11/11) |
| | Reference-embedded SVs | 5 | 100% (5/5) |
| NA12878 | SNV | 2,329 | 99.1% (2, 308/2, 329) |
| | Indel* | 36 | 97.2% (35/36) |
| | SV | 3 | 100% (3/3) |
| | Reference-embedded SVs | 2 | 100% (2/2) |

*Indels greater than 2 bp.

the functional/ORF IGHV genes in this region. The novel SV haplotype in NA19240 contains 3/4 segmental duplication blocks, only lacking the genes *IGHV4-30-4* and *IGHV3-30-5*. Both of these novel SVs are supported by fosmid clones and parental data (**Supplementary Figure S11**).

## Identifying False-Negative and -Positive IGH Variants in Public Datasets

Pitfalls of using short-read data for IGH variant detection and gene annotation have been discussed previously (3, 59). We assessed potential advantages of our approach compared to other high-throughput alternatives. In the CHM1 dataset, we defined a ground truth IGH SNV dataset by directly aligning the IGH locus haplotype from GRCh38 (2) to that of GRCh37 (1). We identified 2,940 SNVs between the two haplotypes in regions of overlap (i.e., non-SV regions). We next aligned Illumina paired-end sequencing data generated from CHM1 (60) and our CHM1 IGenotyper assemblies to GRCh37. We detected 4,433 IGH SNVs in the Illumina dataset, and 2,958 SNVs in the IGenotyper assembly. The Illumina call set included only 73.2% (2,153) of the ground truth SNVs, as well as an additional 2,274 false-positive SNVs (**Figure 4A**). In contrast, the IGenotyper call set included 99.0% (2,912) of the ground truth SNVs, and only 46 (1.6%) false-positive SNVs were called (**Figure 4A**).

We next compared IGenotyper genotypes for NA19240 and NA12878 to 1KGP short-read and microarray data (**Figure 4B**). IGenotyper SNVs were lifted over to GRCh37 ($n$ = 4,474, NA19240; $n$ = 2,868, NA12878), excluding SNVs within SV regions not present in GRCh37 ($n$ = 703, NA19240; $n$ = 737, NA12878) **Supplementary Table S17**). In total, only 57.6% (2,578/4,474) and 76.4% (2,190/2,868) of the IGenotyper SNVs were present in the 1KGP call set for NA19240 and NA12878 (**Figure 4B** and **Supplementary Table S18**). Critically, because insertion SVs are not present in GRCh37, the additional SV-associated SNVs were also missed. Thus, in total, 50.2% (2,599/5,177) and 39.3% (1,415/3,605) of IGenotyper SNVs were absent from 1KGP (**Figure 4B**), including SNVs within 18 and 6 IGHV genes; 1,350/4,474 (NA19240) and 526/2,868 (NA12878) IGenotyper SNVs were not found in any 1KGP sample. The 1KGP call set also included an additional 542 (17.4%) and 76 (3.4%) SNVs (putative false-positives) for NA19240 and NA12878, respectively, including putative false-positive SNVs in 6 and 3 IGHV genes (**Figure 4B**). In contrast to SNVs found only in the 1KGP datasets, we found that in both probands >90% of SNVs detected only by IGenotyper were within short read-inaccessible regions (NA19240, 91.3%, 1,731/1,896; NA12878, 97.1%, 658/678; **Figure 4C** and **Supplementary Table S19**), suggesting that IGenotyper offers improvements in regions that are inherently problematic for short reads. We additionally assessed HWE at the interrogated SNVs, as deviation from HWE is often used to assess SNV quality. In both probands, we found that a greater proportion of SNVs unique to the 1KGP callset deviated from HWE than those called by IGenotyper (**Supplementary Figure S14**).

Finally, we compared larger variants, indels and SVs identified by IGenotyper to those detected by the Human Genome Structural Variation Consortium (HGSV) in NA19240 (34). First, we assessed support for six large known identified SVs in NA19240 (**Figure 3A** and **Supplementary Tables S11 and S21**). BioNano optical mapping data detected events in five out of the six SV regions. The complex SV spanning *IGHV1-8/3-9/IGHV3-64D/5-10-1* was not detected, likely because this event involves a swap of sequences of similar size (∼38 Kb) (2), making it difficult to identify using BioNano. A critical difference to BioNano is that IGenotyper provides nucleotide level resolution allowing for fuller characterization of SV sequence content, including SNVs within these regions (as noted above). In addition to these large SVs, IGenotyper also identified 39 indels (3–49 bps) and an additional 11 SVs (>49 bps; 57–428 bps) in NA19240. Of these, 20 indels and 9 SVs were present in the HGSV integration call set.

## Effects of False-Positive and -Negative Variants on Imputation Accuracy

We explored the potential advantage of our genotyping approach compared to array genotyping and imputation. We applied our long-read capture method to a sample selected from a recent rheumatic heart disease (RHD) GWAS (24), which identified IGH as the primary risk locus. Direct genotyping in this sample was carried out previously using the HumanCore-24 BeadChip ($n$ = 14 SNVs) and targeted Sanger sequencing ($n$ = 8 SNVs); genotypes at additional SNVs were imputed with IMPUTE2 (61), using a combination of 1KGP and population-specific sequencing data as a reference set. We compared IGenotyper SNVs from this sample to directly genotyped variants and imputed variants selected at three hard call thresholds (0.01, 0.05, and 0.1; **Figure 5**). The majority of directly genotyped SNVs (array, 13/14; Sanger, 8/8) were validated by IGenotyper. However, the validation rate varied considerably for the imputed SNVs, depending on the hard call threshold used (**Figure 5A**; 93.8%, 0.01; 92.7%, 0.05; 40.5%, 0.1), with the signal to noise ratio decreasing significantly from the 0.01 to the 0.1 threshold (**Figure 5B**). In all cases, IGenotyper called a substantial number of additional SNVs (**Figure 5A**); the majority of these were located in the proximal region of the locus, which was poorly represented by both directly genotyped and imputed SNVs (**Figure 5C**).

## Sample Multiplexing Leads to Reproducible Assemblies and Variant Calls

An advantage to the capture-based approach employed here is the ability to multiplex samples. To demonstrate this, eight technical replicates of NA12878 were captured, barcoded, pooled, and sequenced on a single Sequel SMRT cell 1M (**Supplementary Table S22**), yielding a mean CCS coverage ranging from 41.3 to 101.6x for each library (**Figure 6A**). We then simulated different plexes (2-, 4-, 16-, 24-, 40-plex) by either combining or partitioning data from this 8 plex, allowing us to assess the impacts of read depth on IGH locus coverage, assembly accuracy, and variant calling. The mean CCS coverage per plex

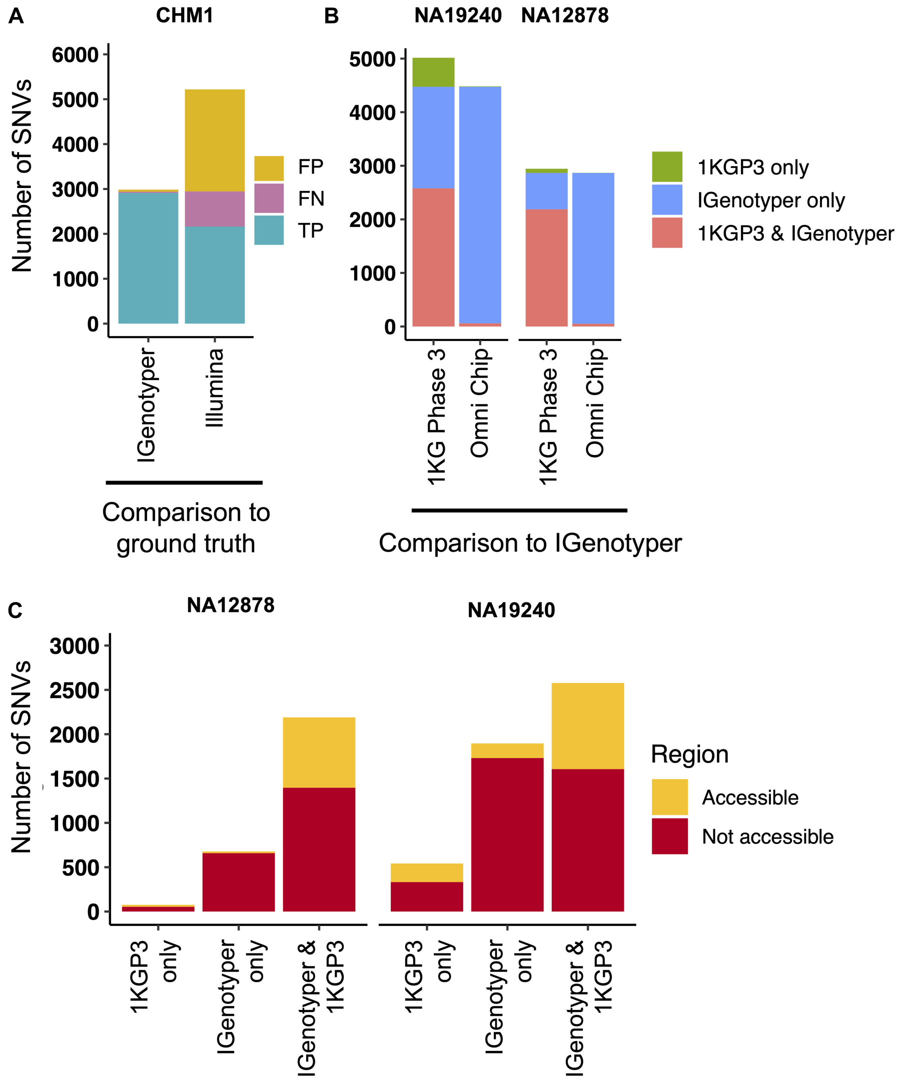


**FIGURE 4 |** Comparison of SNVs identified by IGenotyper to SNVs called using short-read and microarray data. **(A)** SNVs detected by IGenotyper and Illumina/GATK in CHM1 were compared to a CHM1 ground truth SNV dataset; numbers of false-negative, false-positive, and true-positive SNVs in each callset are shown. **(B)** SNVs in the 1KGP Phase 3 (1KGP3) datasets were compared to SNVs detected by IGenotyper in NA19240 and NA12878. The total number of SNVs in each bar sums to the number of overlapping SNVs and the number of SNVs unique to each dataset. **(C)** SNVs found by IGenotyper and the 1KGP dataset, found only by IGenotyper and found only in the 1KGP dataset were partitioned into regions identified as accessible by the 1KGP accessible genome browser track.

ranged from 308.7X (2-plex) to 15.5X (40-plex) (**Figure 6A**). To compare IGenotyper metrics across plexes, we chose the 2-plex sample with the highest CCS coverage to use as the ground-truth dataset; all other assemblies and variant call sets were compared to this sample. The lower coverage 2-plex assembly covered 99.75% of the ground truth assembly with a sequence identity concordance of 99.99%. For the remaining comparisons, mean assembly coverage and sequence concordance estimates ranged from 99.27% (4-plex) to 86.95% (40-plex), and 99.99% (4-plex) to 99.99% (40-plex). The mean number of observed SNVs ranged from 2,471 (2-plex) to 1,936 (40-plex) (**Figure 6B**). When comparing these to ground truth SNVs, we observed high recall rates (>80%), even among the 40-plex assemblies (**Figure 6C**); recall was >90% for all but one of the 4- and 8-plex assemblies (**Figure 6C**). Importantly, although the recall rate of true-positive SNPs decreased as expected in higher plexes, we observed very little variation in the false-positive rate (**Figure 6C**).

## DISCUSSION

For decades, comprehensive genetic analysis of the human IGH locus has been intractable (3, 23, 59). As a result, our understanding of the extent of IGH germline diversity in human populations, and how this diversity contributes to B cell mediated immunity remains incomplete (3, 23). Here, we have leveraged existing genomic haplotype data to design a novel IGH assembly and genotyping framework

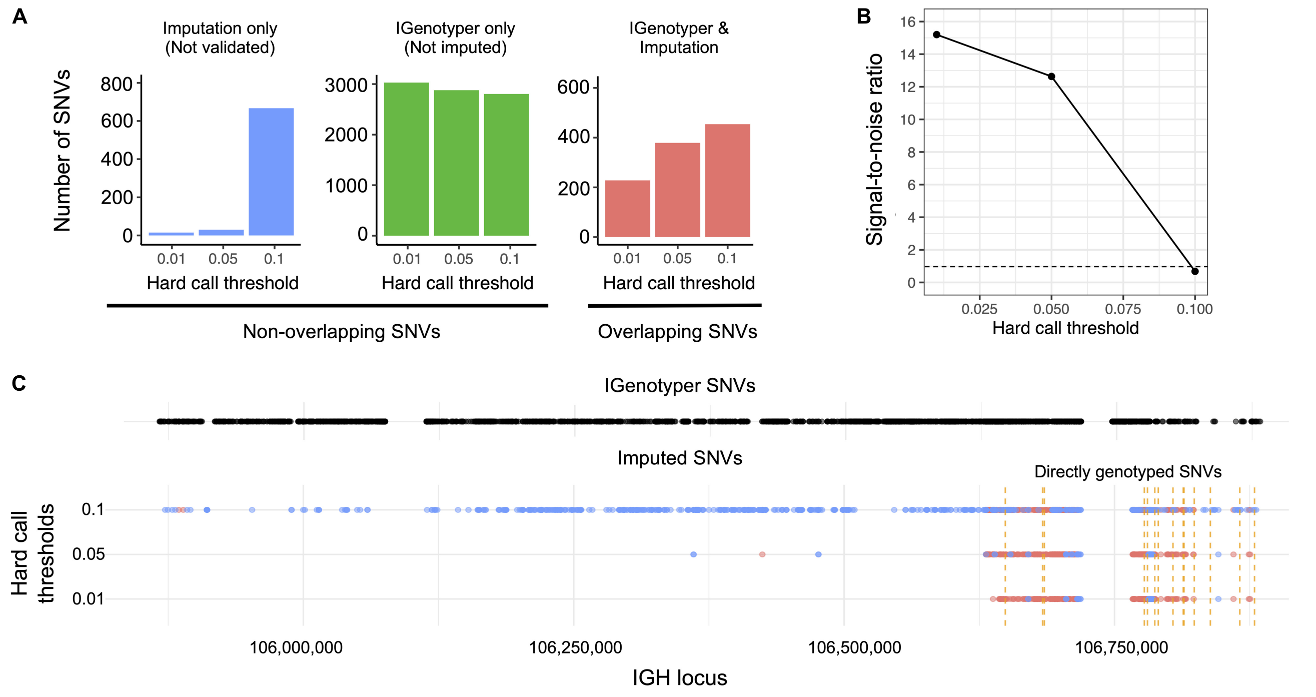


**FIGURE 5 |** Comparison of IGenotyper variant calls to microarray-/imputation-based genotyping. **(A)** SNVs detected by IGenotyper were compared to array-based and imputed SNVs in a sample from a recent RHD GWAS. Imputed SNVs were filtered based on different hard call thresholds (0.01, 0.05, and 0.1). Each set of filtered SNVs was compared to IGenotyper SNVs. The number of imputed SNVs not identified by IGenotyper (blue), SNVs only detected by IGenotyper (green), and overlapping SNVs (red) are shown above. **(B)** The signal-to-noise ratio was calculated using the overlapping SNVs between IGenotyper and imputation, and the SNVs identified only by imputation (not called by IGenotyper). Below the signal-to-noise ratio of 1 (dotted line), there is more noise than signal. **(C)** Positions of SNVs detected by IGenotyper and at different hard call thresholds. Each point represents a SNV. Blue SNVs are SNVs only identified by imputation and not detected by IGenotyper, and red SNVs are SNVs found both by imputation and IGenotyper.

that combines targeted long-read sequencing with a novel bioinformatics toolkit (IGenotyper). This end-to-end pipeline can reconstruct completely phased assemblies via the integration of long-range phase information. Utilizing high-quality CCS reads and derived assemblies facilitates characterization of IGH gene segments and all forms of coding and non-coding variants, including the discovery of novel variants and IG alleles.

We validated our pipeline on eight ethnically diverse human samples with orthogonal data that highlighted multiple strengths of our approach. First, we chose individuals with available BAC and fosmid clone-based assembly datasets for direct comparisons to capture/IGenotyper assemblies, as well as Illumina short-read data for assessing assembly error correction metrics. These comparisons revealed high concordance between assemblies (>99%) in both haploid and diploid samples. Second, we directly validated variants and alleles using trio data. Finally, comparisons to additional variant call sets (1KGP and HGSV) allowed us to assess concordance in variant detection (including indels and SVs), and demonstrate advantages over alternative high-throughput methods. Specifically, compared to microarray-based and short-read sequencing methods, IGenotyper variant call sets were more comprehensive, exhibited greater locus coverage, and were more accurate.

Much recent effort has focused on identifying IG genes/alleles absent from existing databases (2, 5–8, 11, 19, 62–65), revealing many undiscovered alleles in the human population. In our analysis, we identified 15 novel alleles from only two samples. Consistent with previous suggestions of undersampling in non-Caucasian populations (7), the majority ($n$ = 12) were in the Yoruban individual, for which 6 additional novel alleles had been reported in an earlier study (1, 2). Notably, the novel alleles described in NA19240 represent the largest contribution to the IMGT database from a single individual. Related to this point, we also observed high numbers of false-positive/negative IGHV SNVs in 1KGP datasets, reinforcing that efforts to identify IG alleles from 1KGP data should be done with extreme caution (59, 66, 67). An added advantage of our approach is the ability to capture variation outside of IG coding segments and more fully characterize SVs. Although several studies have begun to demonstrate the extent of SV haplotype variation (2, 17, 18, 65), information on polymorphisms within these SVs, and within IGH regulatory and intergenic space remains sparse (23). It is worth noting that, in the few samples analyzed here, the majority of variants were detected in non-coding regions, including SNVs within RS, leader, intronic, and intergenic sequences. We also showed that IGenotyper resolved novel SVs within the complex *IGHV3-30* gene region in both 1KGP diploid samples.

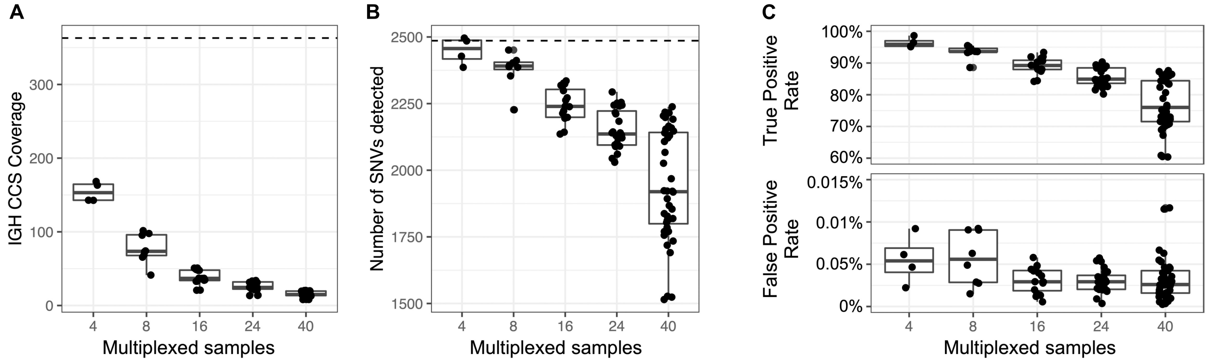


**FIGURE 6 |** Assessment of sample multiplexing on assembly coverage and variant calling. Eight replicates of NA12878 were multiplexed and sequenced on a single Sequel SMRT cell. The eight replicates were combined or down sampled to simulate a sequencing run with 2, 4, 16, 24, and 40 samples. The simulated sample from the 2-plex run with the highest coverage was treated as the ground truth. **(A)** CCS coverage of sequencing runs with different numbers of multiplexed samples. The dotted line represents the coverage of a sample from a 2-plex sequencing run. **(B)** Number of SNVs found across samples per each multiplexed sequencing run. The dotted line represents the number of SNVs detected in a sample from a 2-plex sequencing run. **(C)** True and false positive rate of SNVs of each sample in each multiplexed sequencing run. The SNVs from each sample were compared to SNVs from a sample sequenced in a 2-plex run.

Together, these examples are testament to the fact that our approach represents a powerful tool for characterizing novel IGH variation.

Despite evidence that IG polymorphism impacts inter-individual variation in the antibody response (13, 20, 21, 27), the role of germline variation in antibody function and disease has not been thoroughly investigated. The population-scale IGH screening that will be enabled by this approach will be critical for conducting eQTL studies and integrating additional functional genomic data types to better resolve mechanisms underlying IG locus regulation, which have only so far been applied effectively in model organisms (68–71). Delineating these connections between IGH polymorphism and Ab regulation and function will be critical for understanding genetic contributions to Ab mediated clinical phenotypes (23).

To date, few diseases have been robustly associated to IGH (24, 25, 72, 73). We previously suggested this was due to sparse locus coverage of genotyping arrays and an inability of array SNPs to tag functional IGH variants (2, 3). We have provided further support for this idea here. First, the identification of both putative false negative and positive SNVs in 1KGP samples highlights potential issues with imputation-based approaches using 1KGP samples as a reference set. Second, our direct analysis of capture/IGenotyper data in a sample from a recently conducted GWAS (24) also demonstrated that IGenotyper resulted in a larger set of genotypes, with improved locus coverage compared to imputation. Together, these analyses highlight the potential for our framework to supplement GWASs for both discovery and fine mapping efforts, and through building more robust imputation panels. A strength of our approach is that the user can determine the sequencing depth and locus coverage, depending on whether the intent is to conduct full-locus assemblies or genotyping screens; although the number of detected variants decreases with increased multiplexing, false-positive rates remain low. The recently released, higher throughput Sequel II platform, in combination with read length improvements, will allow for expanded processing of larger cohorts at lower cost.

A current technical limitation of our framework is the decreased efficiency of probes in particular regions of IGH. However, we showed that these regions represent a small fraction of IGH, with overall negligible impacts on locus coverage and assembly quality. Future iterations of target-enrichment protocols will improve efficiency through methodological or reagent modifications. A key strength of IGenotyper is its flexibility to accommodate other data types; e.g., users interested in complete haplotype characterization can already provide long-range phase information to inform a complete diploid assembly. We envision other forms of data will be leveraged in future applications. A second limitation is the potential to miss unknown sequences not specifically targeted by capture probes. We expect this issue to be mitigated in the future as more IG haplotypes are sequenced using the method described here, as well as through the application of large-insert clone assembly and WGS approaches; enumeration of these data will allow for refinement on protocol design and IGenotyper functionality. Ultimately, we promote an advance toward a more extensive collection of IGH haplotype reference datasets and variants as a means to leverage more sophisticated strategies for variant calling; for example, population reference graph approaches, which have been shown to be effective in other hyperpolymorphic immune loci (74). Looking forward, we expect our approach will lead to more comprehensive datasets that will augment and extend existing IG germline databases, such as IMGT (4) and VDJbase (75), and facilitate more effective modes of sharing IG polymorphism and haplotype data.

To the best of our knowledge, this is the first combined molecular protocol and analytical pipeline that can provide comprehensive genotype and annotation information across the

IGH locus, with the added ability to be applied to a large number of samples in a high-throughput manner. Given the importance of antibody repertoire profiling in health and disease, characterizing germline variation in the IG regions will continue to become increasingly important. Our strategy moves toward the complete ascertainment of IG germline variation, a prerequisite for understanding the genetic basis of Ab-mediated processes in human disease.

## DATA AVAILABILITY STATEMENT

All raw data and fosmid assemblies used to validate our approach can be found under BioProject PRJNA555323.

## ETHICS STATEMENT

Ethical approval was granted previously for the usage of the RHD GWAS study sample by the Hospital Ethics Committee at the Hôpital de Gaston-Bourret and the Comité d'Evaluation Ethique de l'Inserm as well as the Oxford University Tropical Research Ethics Committee. Previous consent was granted for the hydatidiform mole sample procured from the laboratory of Dr. Urvashi Surti. The sample was obtained from the participant with written informed consent to be used for research purposes.

## AUTHOR CONTRIBUTIONS

OLR, WSG, AJS, MLS, AB, and CTW conceived and planned the study. OLR and TP performed the computational analyses. OLR and CTW wrote the manuscript with contributions from TP, EEE, AJS, WSG, MLS, and AB. OLR wrote the code. EEE, KA, and TP provided the additional samples. WSG, ME, and MLS prepared the sequencing libraries. GD, JP, MLS, and RS performed the sequencing. CTW, AB, and MLS supervised the experiments, analysis, and data interpretation. All authors read and approved the final manuscript.

## FUNDING

This work was supported, in part, by grants from the U.S. National Institutes of Health R21AI142590 (to CTW and WAM), NIH R24AI138963 (to CTW and MLS), NIH R21AI117407 (to AB and AJS), NIH 1F31NS108797 (to OLR), NIH HG010169 (to EEE), British Heart Foundation PG/14/26/30509 (to TP), and Medical Research Council UK Fellowship G1100449 (to TP). EEE is an investigator of the Howard Hughes Medical Institute.

## ACKNOWLEDGMENTS

This study makes use of a sample from an individual recruited by the Pacific Islands Rheumatic Heart Disease Genetics Network. This work was supported in part through the computational resources and staff expertise provided by the Scientific Computing at the Icahn School of Medicine at Mount Sinai. This manuscript has been released as a pre-print at bioRxiv (76).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.02136/full#supplementary-material

## REFERENCES

1. Matsuda F, Ishii K, Bourvagnet P, Kuma K-I, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med.* (1998) 188:2151–62. doi: 10.1084/jem.188.11.2151
2. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J. Complete haplotype sequence of the human immunoglobulin heavy-chain variable,diversity,and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet.* (2013) 92:530–46. doi: 10.1016/j.ajhg.2013.03.004
3. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation,missing data,and implications for human disease. *Genes Immun.* (2012) 13:363–73. doi: 10.1038/gene.2012.12
4. Lefranc M-P, Lefranc M-P. IMGT,the international ImMunoGeneTics database. *Nucleic Acids Res.* (2001) 29:207–9. doi: 10.1093/nar/29.1.207
5. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol.* (2010) 184:6986–92. doi: 10.4049/jimmunol.1000445
6. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad. Sci USA.* (2015) 112:E862–70. doi: 10.1073/pnas.1417683112
7. Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline ig gene repertoire. *J Immunol.* (2015) 194:4371–8. doi: 10.4049/jimmunol.1500118
8. Corcoran MM, Phad GE, Vázquez Bernat N, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun.* (2016) 7:13642. doi: 10.1038/ncomms13642
9. Thörnqvist L, Ohlin M. The functional 3'-end of immunoglobulin heavy chain variable (IGHV) genes. *Mol Immunol.* (2018) 96:61–8. doi: 10.1016/j.molimm.2018.02.013
10. Calonga-Solís V, Malheiros D, Beltrame MH, De Brito Vargas L, et al. Unveiling the diversity of immunoglobulin heavy constant gamma (IGHG) gene segments in brazilian populations reveals 28 novel alleles and evidence of gene conversion and natural selection. *Front Immunol.* (2019) 10:1161. doi: 10.3389/fimmu.2019.01161
11. Wang Y, Jackson KJL, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol.* (2008) 86:111–5. doi: 10.1038/sj.icb.7100144
12. Milner EC, Hufnagle WO, Glas AM, Suzuki I, Alexander C. Polymorphism and utilization of human VH Genes. *Ann N Y Acad Sci.* (1995) 764:50–61. doi: 10.1111/j.1749-6632.1995.tb55806.x
13. Sasso EH, Johnson T, Kipps TJ. Expression of the immunoglobulin VH gene 51p1 is proportional to its germline gene copy number. *J Clin Invest.* (1996) 97:2074–80. doi: 10.1172/JCI118644
14. Chimge N-O, Pramanik S, Hu G, Lin Y, Gao R, Shen L, et al. Determination of gene organization in the human IGHV region on single chromosomes. *Genes Immun.* (2005) 6:186–93. doi: 10.1038/sj.gene.6364176

15. Pramanik S, Cui X, Wang H-Y, Chimge N-O, Hu G, Shen L, et al. Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. *BMC Genomics.* (2011) 12:78. doi: 10.1186/1471-2164-12-78
16. Kidd MJ, Jackson KJL, Boyd SD, Collins AM. DJ Pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. *J Immunol.* (2016) 196:1158–64. doi: 10.4049/jimmunol.1501401
17. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, et al. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol.* (2012) 188:1333–40. doi: 10.4049/jimmunol.1102097
18. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat Commun.* (2019) 10:628. doi: 10.1038/s41467-019-08489-3
19. Luo S, Yu JA, Li H, Song YS. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *Life Sci Alliance.* (2019) 2:e201800221. doi: 10.26508/lsa.201800221
20. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires,correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep.* (2016) 6:20842. doi: 10.1038/srep20842
21. Glanville J, Kuo TC, von Büdingen H-C, Guey L, Berka J, Sundar PD, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci USA.* (2011) 108:20066–71. doi: 10.1073/pnas.1107498108
22. Rubelt F, Bolen CR, McGuire HM, Vander Heiden JA, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun.* (2016) 7:11112. doi: 10.1038/ncomms11112
23. Watson CT, Glanville J, Marasco WA. The individual and population genetics of antibody immunity. *Trends Immunol.* (2017) 38:459–70. doi: 10.1016/j.it.2017.04.003
24. Parks T, Mirabel MM, Kado J, Auckland K, Nowak J, Rautanen A, et al. Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat Commun.* (2017) 8:14946. doi: 10.1038/ncomms14946
25. Witoelar A, Rongve A, Almdahl IS, Ulstein ID, Engvig A, White LR, et al. Meta-analysis of Alzheimer's disease on 9, 751 samples from Norway and IGAP study identifies four risk loci. *Sci Rep.* (2018) 8:18088. doi: 10.1038/s41598-018-36429-6
26. Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell.* (2010) 143:837–47. doi: 10.1016/j.cell.2010.10.027
27. Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G. A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to *haemophilus influenzae* type b disease. *J Clin Invest.* (1996) 97:2277–82. doi: 10.1172/JCI118669
28. Luo S, Yu JA, Song YS. Estimating copy number and allelic variation at the immunoglobulin heavy chain locus using short reads. *PLoS Comput Biol.* (2016) 12:e1005117. doi: 10.1371/journal.pcbi.1005117
29. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. *Am J Hum Genet.* (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023
30. Neville MJ, Lee W, Humburg P, Wong D, Barnardo M, Karpe F, et al. High resolution HLA haplotyping by imputation for a British population bioresource. *Hum Immunol.* (2017) 78:242–51. doi: 10.1016/j.humimm.2017.01.006
31. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun.* (2017) 8:1326. doi: 10.1038/s41467-017-01343-4
32. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature.* (2015):608–11. doi: 10.1038/nature13907
33. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. *Cell.* (2019) 176:663–75.e19. doi: 10.1016/j.cell.2018.12.019
34. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun.* (2019) 10:1784.
35. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* (2017) 27:677–85. doi: 10.1101/gr.214007.116
36. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun.* (2017) 18:127–34. doi: 10.1038/gene.2017.10
37. Suzuki S, Ranade S, Osaki K, Ito S, Shigenari A, Ohnuki Y, et al. Reference grade characterization of polymorphisms in full-length HLA class I and II genes with short-read sequencing on the ION PGM system and long-reads generated by single molecule,real-time sequencing on the pacbio platform. *Front Immunol.* (2018) 9:2294. doi: 10.3389/fimmu.2018.02294
38. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* (2019) 37:1155–62. doi: 10.1038/s41587-019-0217-9
39. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods.* (2015) 12:780–6. doi: 10.1038/nmeth.3454
40. Hafford-Tear NJ, Tsai Y-C, Sadan AN, Sanchez-Pintado B, Zarouchlioti C, Maher GJ, et al. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. *Genet Med.* (2019) 21:2092–102. doi: 10.1038/s41436-019-0453-x
41. Ebbert MTW, Farrugia SL, Sens JP, Jansen-West K, Gendron TF, Prudencio M, et al. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol Neurodegener.* (2018) 13:46. doi: 10.1186/s13024-018-0274-4
42. Hoff SNK, Baalsrud HT, Tooming-Klunderud A, Skage M, Richmond T, Obernosterer G, et al. Long-read sequence capture of the haemoglobin gene clusters across codfish species. *Mol Ecol Resour.* (2019) 19:245–59. doi: 10.1111/1755-0998.12955
43. Bethune K, Mariac C, Couderc M, Scarcelli N, Santoni S, Ardisson M, et al. Long−fragment targeted capture for long−read sequencing of plastomes. *Appl Plant Sci.* (2019) 7:e1243. doi: 10.1002/aps3.1243
44. Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, Midwinter W, et al. HLA typing for the next generation. *PLoS One.* (2015) 10:e0127153. doi: 10.1371/journal.pone.0127153
45. Bultitude WP, Gymer AW, Robinson J, Mayor NP, Marsh SGE. KIR2DL1 allele sequence extensions and discovery of 2DL1*0010102 and 2DL1*0010103 alleles by DNA sequencing. *Hladnikia.* (2018) 91:546–7. doi: 10.1111/tan.13269
46. Turner TR, Hayhurst JD, Hayward DR, Bultitude WP, Barker DJ, Robinson J, et al. Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines. *Hladnikia.* (2018) 91:88–101. doi: 10.1111/tan.13184
47. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* (2011) 29:24–6. doi: 10.1038/nbt.1754
48. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics.* (2012) 13:238. doi: 10.1186/1471-2105-13-238
49. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, et al. WhatsHap: fast and accurate read-based phasing. *bioRxiv.* (2016).doi: 10.1101/085050 [Preprint].
50. Rodriguez OL, Ritz A, Sharp AJ, Bashir A. MsPAC: a tool for haplotype-phased structural variant detection. *Bioinformatics.* (2019) 36:922–4. doi: 10.1093/bioinformatics/btz618

51. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* (2017) 27:722–36. doi: 10.1101/gr.215087.116
52. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* (2011) 43:491–8. doi: 10.1038/ng.806
53. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* (2015) 526:68–74. doi: 10.1038/nature15393
54. Samorodnitsky E, Datta J, Jewell BM, Hagopian R, Miya J, Wing MR, et al. Comparison of custom capture for targeted next-generation DNA sequencing. *J Mol Diagn.* (2015) 17:64–75. doi: 10.1016/j.jmoldx.2014.09.009
55. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol.* (2018) 36:1174–82. doi: 10.1038/nbt.4277
56. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods.* (2016) 13:1050–4. doi: 10.1038/nmeth.4035
57. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome Res.* (2017) 27:757–67. doi: 10.1101/gr.214874.116
58. Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet.* (2015) 16:627–40. doi: 10.1038/nrg3933
59. Watson CT, Matsen FA, Jackson KJL, Bashir A, Smith ML, Glanville J, et al. Comment on 'A database of human immune receptor alleles recovered from population sequencing data'. *J Immunol.* (2017) 198:3371–3. doi: 10.4049/jimmunol.1700306
60. Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, et al. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* (2014) 24:2066–76. doi: 10.1101/gr.180893.114
61. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* (2009) 5:e1000529. doi: 10.1371/journal.pgen.1000529
62. Gadala-Maria D, Gidoni M, Marquez S, Vander HJA, Kos JT, Watson CT, et al. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Front Immunol.* (2019) 10:129. doi: 10.3389/fimmu.2019.00129
63. Zhang W, Wang I-M, Wang C, Lin L, Chai X, Wu J, et al. IMPre: an accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol.* (2016) 7:457. doi: 10.3389/fimmu.2016.00457
64. Ralph DK, Matsen F. A 4th. Consistency of VDJ Rearrangement and substitution parameters enables accurate b cell receptor sequence annotation. *PLoS Comput Biol.* (2016) 12:e1004409. doi: 10.1371/journal.pcbi.1004409
65. Kirik U, Persson H, Levander F, Greiff L, Ohlin M. Antibody heavy chain variable domains of different germline gene origins diversify through different paths. *Front Immunol.* (2017) 8:1433. doi: 10.3389/fimmu.2017.01433
66. Khatri I, Berkowska MA, van den Akker EB, Teodosio C, Reinders MJT, van Dongen JJM. Population matched (PM) germline allelic variants of immunoglobulin (IG) loci: new pmIG database to better understand IG repertoire and selection processes in disease and vaccination. *bioRxiv.* (2020). doi: 10.1101/2020.04.09.033530 [Preprint].
67. Yu Y, Ceredig R, Seoighe C. A database of human immune receptor alleles recovered from population sequencing data. *J Immunol.* (2017) 198:2202–10. doi: 10.4049/jimmunol.1601710
68. Choi NM, Loguercio S, Verma-Gaur J, Degner SC, Torkamani A, Su A I, et al. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *J Immunol.* (2013) 191:2393–402. doi: 10.4049/jimmunol.1301279
69. Kumar S, Wuerffel R, Achour I, Lajoie B, Sen R, Dekker J, et al. Flexible ordering of antibody class switch and V(D)J joining during B-cell ontogeny. *Genes Dev.* (2013) 27:2439–44. doi: 10.1101/gad.227165.113
70. Feldman S, Wuerffel R, Achour I, Wang L, Carpenter PB, Kenter AL. 53BP1 contributes to Igh locus chromatin topology during class switch recombination. *J Immunol.* (2017) 198:2434–44. doi: 10.4049/jimmunol.1601947
71. Barajas-Mora EM, Feeney AJ. Enhancers as regulators of antigen receptor loci three-dimensional chromatin structure. *Transcription* (2020) 11:37–51. doi: 10.1080/21541264.2019.1699383
72. Tsai F-J, Lee Y-C, Chang J-S, Huang L-M, Huang F-Y, Chiu N-C, et al. Identification of novel susceptibility loci for kawasaki disease in a han chinese population by a genome-wide association study. *PLoS One.* (2011) 6:e16853. doi: 10.1371/journal.pone.0016853
73. Dhande IS, Kneedler SC, Joshi AS, Zhu Y, Hicks MJ, Wenderfer SE, et al. Germ-line genetic variation in the immunoglobulin heavy chain creates stroke susceptibility in the spontaneously hypertensive rat. *Physiol Genomics.* (2019) 51:578–85. doi: 10.1152/physiolgenomics.00054.2019
74. Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. *Nat Genet.* (2015) 47:682–8. doi: 10.1038/ng.3257
75. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, et al. VDJbase: an adaptive immune receptor genotype and haplotype database. *Nucleic Acids Res.* (2020) 48:D1051–6. doi: 10.1093/nar/gkz872
76. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *bioRxiv.* (2020). doi: 10.1101/2020.04.19.049270 [Preprint].

**Conflict of Interest:** EE is on the Scientific Advisory Board (SAB) of DNAnexus, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer GY declared a past co-authorship with one of the authors, CW, to the handling editor.

ORIGINAL RESEARCH
published: 09 October 2020
doi: 10.3389/fimmu.2020.582927

# Efficient Sequencing, Assembly, and Annotation of Human KIR Haplotypes

*David Roe*[1]*, *Jonathan Williams*[2], *Keyton Ivery*[2], *Jenny Brouckaert*[2], *Nick Downey*[3], *Chad Locklear*[3], *Rui Kuang*[1,4] *and Martin Maiers*[5]

[1] Bioinformatics and Computational Biology, University of Minnesota, Rochester, MN, United States, [2] DNA Identification Testing Division, Laboratory Corporation of America Holdings, Burlington, NC, United States, [3] Integrated DNA Technologies, Inc., Coralville, IA, United States, [4] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, United States, [5] Center for International Blood and Marrow Transplant Research, Minneapolis, MN, United States

The homology, recombination, variation, and repetitive elements in the natural killer-cell immunoglobulin-like receptor (KIR) region has made full haplotype DNA interpretation impossible in a high-throughput workflow. Here, we present a new approach using long-read sequencing to efficiently capture, sequence, and assemble diploid human KIR haplotypes. Probes were designed to capture KIR fragments efficiently by leveraging the repeating homology of the region. IDT xGen® Lockdown probes were used to capture 2–8 kb of sheared DNA fragments followed by sequencing on a PacBio Sequel. The sequences were error corrected, binned, and then assembled using the Canu assembler. The location of genes and their exon/intron boundaries are included in the workflow. The assembly and annotation was evaluated on 16 individuals (8 African American and 8 Europeans) from whom ground truth was known *via* long-range sequencing with fosmid library preparation. Using only 18 capture probes, the results show that the assemblies cover 97% of the GenBank reference, are 99.97% concordant, and it takes only 1.8 haplotigs to cover 75% of the reference. We also report the first assembly of diploid KIR haplotypes from long-read WGS. Our targeted hybridization probe capture and sequencing approach is the first of its kind to fully sequence and phase all diploid human KIR haplotypes, and it is efficient enough for population-scale studies and clinical use. The open and free software is available at https://github.com/droeatumn/kass and supported by a environment at https://hub.docker.com/repository/docker/droeatumn/kass.

**Keywords: killer-cell immunoglobulin-like receptor, assembly, DNA, haplotype, annotation, natural killer**

OPEN ACCESS

**Edited by:**
*Lutz Walter,*
*Deutsches Primatenzentrum,*
*Germany*

**Reviewed by:**
*Stephen K. Anderson,*
*National Cancer Institute at Frederick,*
*United States*
*Lisbeth Guethlein,*
*Stanford University, United States*

***Correspondence:**
*David Roe*
*roex0050@umn.edu*

**Specialty section:**
*This article was submitted to*
*NK and Innate*
*Lymphoid Cell Biology,*
*a section of the journal*
*Frontiers in Immunology*

**Received:** *14 July 2020*
**Accepted:** *17 September 2020*
**Published:** *09 October 2020*

**Citation:**
*Roe D, Williams J, Ivery K,*
*Brouckaert J, Downey N, Locklear C,*
*Kuang R and Maiers M (2020) Efficient*
*Sequencing, Assembly, and*
*Annotation of Human KIR Haplotypes.*
*Front. Immunol. 11:582927.*
*doi: 10.3389/fimmu.2020.582927*

## INTRODUCTION

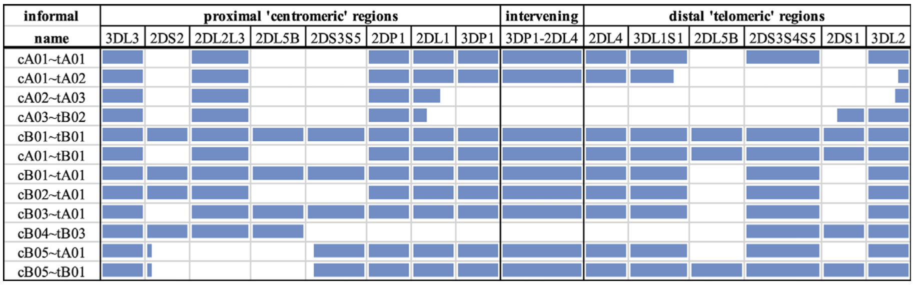
The protein coding killer-cell immunoglobulin-like receptor (KIR) genes span ~10–16 kb each, with pseudogenes that are ~5 and ~13 kb. Alleles from any two genes are over 85–98% identical. Frequent recombination throughout the ~70–270 kb haplotypes has made their order and copy number highly variable. The genes encode proteins that recognize human leukocyte antigen (HLA) and its peptide, and along with other receptors, initiate signaling pathways in natural killer (NK) cells that can lead to the release of cytokines or to the death of the target cell (infected, cancerous, foreign, etc.). Some of these ligand-receptor interactions stimulate the NK cell to react, while some inhibit the NK cell from reacting until the ligand is missing. NKs and their KIR receptors are essential to human health and have functional roles that impact viral infections, pregnancy, autoimmune diseases, transplantation, and immunotherapy (1–7).

Genetic interpretation of exonic-or-lower resolutions from next generation sequencing (NGS) is often ambiguous and unphased, and therefore limits precise understanding of how KIR sequences affect phenotypes. The importance of high-throughput high-resolution typing is exemplified by the fact that the genes contain extensive exonic SNP and short insertion/deletion (indel) variations which rivals that of its binding partner: HLA class I (1, 8). Over 300 full-length DNA and almost 1,000 protein reference alleles have been reported in IPD-KIR (9). All resolutions except haplotyping are ambiguous or require statistical phasing from few references. A cost-effective high-throughput method that could characterize all the sequences within the KIR haplotypes in cis could advance that understanding and clarify previously ambiguous and/or contradictory evidence. To date, the only approach for full haplotyping was to physically separate and amplify maternal and paternal haplotypes *via* fosmids for subsequent sequencing (10–14), a process whose expense has generally prohibited its use in large-scale association studies. While high-resolution haplotyping by fosmid clones or full gene by PCR is costly and inefficient, low resolution genotyping of gene presence/absence or copy number provide limited information for functional analysis and association tests.

Like much of chromosome 19, the KIR region is dense with repetitive elements, which have provided the mechanisms for its recent evolution by tandem duplication and homologous recombination events. Dozens of distinct gene-content haplotypes are seen in Europeans alone (11, 12, 15, 16). Previous reports have documented over ten distinct common haplotype structures (13). **Figure 1** provides an overview of the most common haplotype structures and their informal names. KIR haplotypes are named in two halves: “c” for centromeric (i.e., proximal) and “t” for telomeric (i.e., distal) separated by a recombination hotspot contained in the ~10 kb intergenic region between *KIR3DP1* and *KIR2DL4*. Each half is also labelled “A” or “B”, designating one of two families of haplotypes, based on the gene content (17). The A family haplotype denotes haplotypes with one main gene content structure and relatively large allelic variation. The B family of haplotypes denotes a class of haplotypes with relatively more structural variation and less allelic variation. The haplotype named “cA01~tB01”, for example, means the first (01) centromeric A region in cis (“~”) with the first telomeric B region.

It is difficult to interpret KIR haplotypes for an individual human genome given the reads from high-throughput sequencing when the structural arrangements are unknown. This is largely due to read lengths from prevailing technologies being too short to map unambiguously to the repetitive and homologous KIR genes. Even if the reads could be uniquely placed, they require statistical phasing that is difficult due to lack of phased high-resolution reference libraries. As a consequence, the reads from the KIR region are largely ignored, mis-interpreted, or under-interpreted in current whole genome sequencing (WGS) studies. Therefore, the properties of the KIR region require more careful and specific interpretations than most other regions in human genome.

Here we present an approach that leverages PacBio’s long-read circular consensus sequence (CCS) reads to span DNA homology, and gene homology to efficiently capture 2–8 kb fragments of DNA. It is a workflow to capture, sequence, assemble, and annotate diploid human KIR haplotypes. And it also has broader implications to other genomic regions with variable or repetitive regions alternating with constant regions. When applied to a cohort of 8 African Americans and a cohort of 8 Europeans, the results demonstrate that every KIR gene and intergene contains constant regions that are targetable by capture probes, and that by targeting the constant regions, the variable regions can be captured and sequenced by standard PacBio workflows. Further, maximizing this paradigm shows that 18 short probe sequences can capture KIR haplotypes and allow



| informal | proximal 'centromeric' regions | | | | | | | | intervening | distal 'telomeric' regions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| name | 3DL3 | 2DS2 | 2DL2L3 | 2DL5B | 2DS3S5 | 2DP1 | 2DL1 | 3DP1 | 3DP1-2DL4 | 2DL4 | 3DL1S1 | 2DL5B | 2DS3S4S5 | 2DS1 | 3DL2 |
| cA01~tA01 | ■ | | ■ | | | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | | ■ |
| cA01~tA02 | ■ | | ■ | | | ■ | ■ | ■ | ■ | ■ | ▪ | | | | ▪ |
| cA02~tA03 | ■ | | ■ | | | ■ | ▪ | | | | | | | | ▪ |
| cA03~tB02 | ■ | | ■ | | | ■ | ▪ | | | | | | | ▪ | ■ |
| cB01~tB01 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| cA01~tB01 | ■ | | ■ | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| cB01~tA01 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | | ■ |
| cB02~tA01 | ■ | ■ | ■ | | | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | | ■ |
| cB03~tA01 | ■ | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | | ■ |
| cB04~tB03 | ■ | ■ | ■ | ■ | | | | | | | | | ■ | ■ | ■ |
| cB05~tA01 | ■ | ▪ | | | ▪ | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | | ■ |
| cB05~tB01 | ■ | ▪ | | | ▪ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

**FIGURE 1** | Common KIR haplotype structures and their names. Blue regions represent the presence of a gene or the *KIR3DP1-KIR2DL4* intervening intergenic region. Some genes are partially blue to indicate their portion of a fusion allele. The first column contains the informal name for the haplotype. Each haplotype name is a combination of its two regions: centromeric (proximal) regions are preceded with a “c” and telomeric (distal) regions are preceded with a “t”. Not all haplotype structures are shown.

their unambiguous assembly. Finally, this is an efficient approach that requires no prior knowledge of the individual or references, only utilizes standard lab workflows, and is available in free and open software.

## MATERIALS AND METHODS

### Overview

The goal of the experiment was to create a set of capture probes and a bioinformatics workflow to efficiently assemble full KIR haplotypes from PacBio CCS reads. The experiments to capture, sequence, assemble, and annotate are depicted graphically in **Figure 2**. The major steps consist of

1. Design capture probes.
2. Use the probes to capture the KIR DNA fragments *in vitro* or *in silico* per individual.
3. Sequence the fragments on PacBio Sequel.
4. Error correct the sequences.
5. Bin the sequences per KIR region and gene.
6. *de novo* assemble all the sequences together and each gene bin separately.
7. Annotate the assembled sequences with their genes and exon/intron locations.

### Step 1: Design Capture Probes

Published KIR haplotypes sequences (36 at the time of this study) as well as all allele sequences from IPD-KIR 2.7.1 (18) were used to generate 200 candidate capture probes. The design of the 120-base probes was coordinated with a combination of automated and manual Integrated DNA Technologies (IDT) design tools, including a strain typer alignment tool and a xGen® Lockdown probe design tool. The candidate set of probes was reduced by leveraging sequence homology. First, the haplotype sequences were aligned to two KIR haplotypes (GenBank accessions GU182358 and GU182339). These two reference haplotypes, which together contain all KIR genes, were annotated *via* RepeatMasker (19, 20). The candidates were prioritized by the highest number of times each aligned to the two reference sequences but not to repetitive elements. The set was chosen

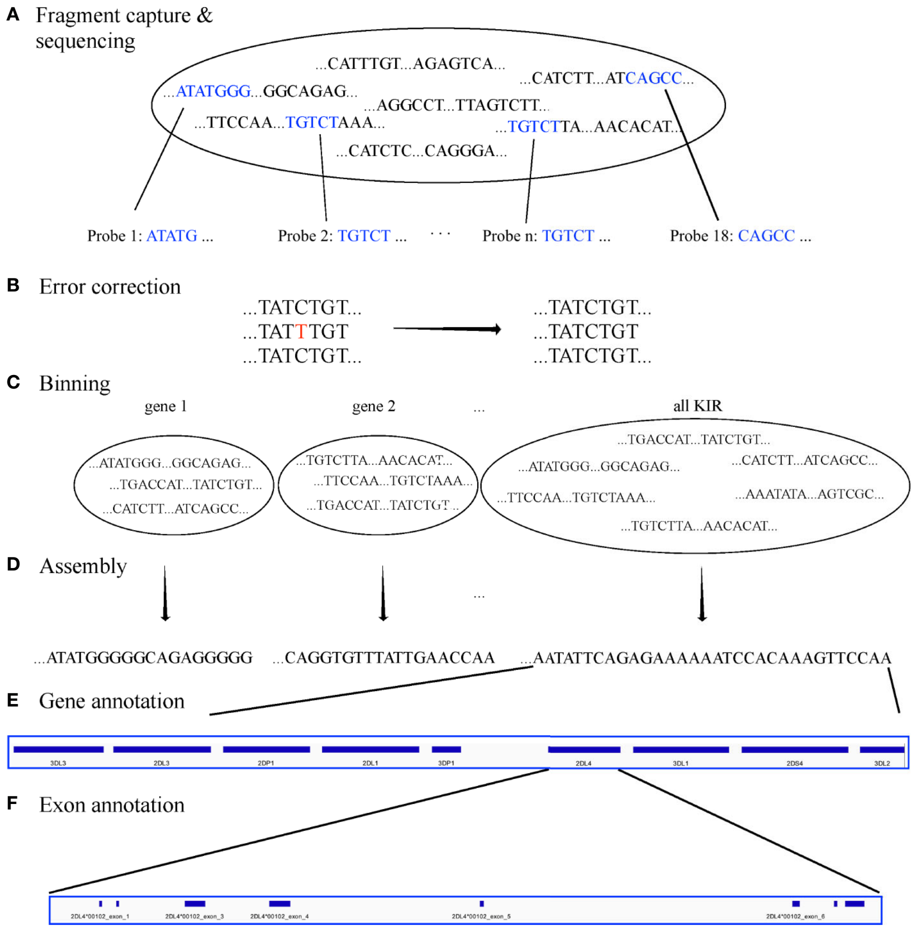


**FIGURE 2** | Workflow. The workflow starts with fragment capture *in vitro* or *in silico* and PacBio assembly **(A)**. The sequences are error corrected **(B)** and binned by KIR region and KIR gene **(C)** before *de novo* assembly of each bin **(D)**. Finally, the assembled haplotigs/haplotypes are annotated by gene **(E)** and exon **(F)**.

by iteratively adding the probe with the highest alignment hit count until both reference haplotypes were covered by less than the expected average DNA fragment length of ~4–5 kb. Ultimately, experiments were conducted on the original set of 200, a minimal set of 15, and a refined set of 18 capture probes.

## Steps 2–3: Capture and Sequence the DNA Fragments per Individual

Targeted hybridization probe capture and sequencing was performed (**Figure 2A**) as described using PacBio unsupported protocol PN101-388-000 (21) with the following modifications. Two µg Human Genomic DNA suspended in 200 µl if Elution Buffer was sheered with Covaris G-tubes to 6–8 kb according to manufactures instructions followed by 1:1 PB AMPure bead cleanup. The target fragment size of 6–8 kb was chosen to maximize the ability for capture and to allow proper phasing and the generation of CCS reads, or a consensus sequence of one captured DNA fragment per sequence well. Individual specimens were then library prepped with the KAPA library prep kit (Roche) which consisted of end repair and ligation of uniquely barcoded adapters that also contained the PacBio Universal Primer sequence. After a 0.8X PB AMPure bead clean up samples were enriched with eight PCR cycles in a 200 µl reaction using LA Taq by Takara and PacBio Universal primers followed by a 1:2 PB AMPure bead cleanup. Sample concentrations were measured on the Promega Quantus and 2 µg was size selected for greater than 2 kb fragments on the Blue Pippin System (Sage Sciences).

Eight multiplexed samples for Sequel sequencing were pooled at this point at 0.25 µg per sample and 1.5 µg of pooled size selected DNA, 5 µl of 1 mg/ml Human Cot-1 DNA by Thermo Fisher, and 10 µl of 100 µM PacBio Universal Primer were dried in a Speed Vac with no heat. IDT xGen® lockdown reagents and probes were used to resuspend the genomic/Cot-1/primer mixture according to manufactures instructions and incubated at 70°C for 4 h followed by IDT xGEN washes®. DNA was then removed from streptavidin beads and further enriched with 15 PCR cycles using LA Taq. DNA fragments were then library prepped and sequenced eight samples per SMRT cell on the Sequel according to PacBio instructions. Raw sequence data was then demultiplexed and CCS reads generated on SmrtLink 6.0 using 99.9% subread accuracy filter for generation.

## Steps 4–6: Correct, Bin, and Assemble the Sequences

For both targeted and WGS, the fastq sequences were error corrected with LoRMA (22). The sequences were binned for on-KIR and also binned per genic or intergenic region in silico (**Figures 2B, C**) using the 18 capture probes for on-off KIR detection and 32,230 gene probes. The gene probes are 25mers and are detailed in a recent manuscript by Roe et al. that has been submitted for peer review and preprinted on bioRxiv (23). Synthetic probe matching was conducted *via* bbduk (24) with parameters "k=25 maskmiddle=f overwrite=t rename=t nzo=t rcomp=t ignorebadquality=t". This effectively removed any off-KIR sequences and binned the sequences into 15 loci: 12 protein-coding genes, 2 pseudo-genes, and the intergenic region between *KIR3DP1* and *KIR2DL4*. Sequences in each bin were *de novo* assembled with Canu 2.0 (25) (with default parameters except "genomeSize=200k") for each bin separately and all KIR sequences together (**Figure 2D**). The assemblies utilized only the captured sequences and were not assisted by any prior information, `including individual genotypes or reference libraries.

## Step 7: Annotate the Assembled Sequences

The capture probes were aligned to the haplotype-specific assembled sequences (i.e. haplotigs), and their patterns allowed gene-specific sequences to be extracted from the haplotypes (**Figure 2E**). The details are presented in the previously-mentioned preprint. At a high level, the algorithm uses the bowtie2 alignment pattern of the 18 capture probes across the haplotigs/haplotypes to define locus-specific features. Within each feature, the locations of the exons, introns, and untranslated elements were located by searching for inter-element boundaries with 16 base sequences as defined by the full haplotype MSA of 68 human haplotypes (**Figure 2F**). Each sequence contains 8 bases from one region and 8 from the other. For example, ACACGTGGGTGAGTCC spans the boundary between *KIR2DL4*'s exon two and intron two; the first eight characters are from the second exon, and the last eight bases from the second intron. Boundary regions are flexible up to 3 mismatches if necessary. The locations of these elements allows for the annotation of protein, cDNA, and full-gene alleles with respect to names assigned in IPD-KIR. The haplotigs were ultimately annotated in GenBank's tbl format. BioJava (26) was used for some of the sequence processing. Reports on the assembly (or the raw sequences) were generated from Minimap2 (27), Qualimap (28), NanoPack (29), QUAST (30), Simple Synteny (31), and Tablet (32).

## Evaluation of the Workflow

The capture-sequence-assemble workflow was evaluated on a cohort of 16 individuals whose haplotypes had previously been sequenced using fosmid separation and long-read sequencing (13, 33). The assembled haplotype-specific sequences (i.e., haplotigs), were evaluated by their phased coverage and concordance with the reference sequences as well as the number of haplotigs it takes to phase 75% of the haplotype (i.e., LG75). The LG metric is a standard metric for assembly evaluation; it is particularly appropriate for variable length haplotypes like KIR to evaluate their haplotigs by the fraction of the haplotype as opposed to number of bases. Of the 16 individuals, 8 are of European (EUR) ancestries (GenBank haplotype sequences KP420437-9, KP420440-6, KU645195-8, and KU842452) and 8 are of African American (AFA) ancestries (GenBank haplotype sequences MN167507, MN167510, MN167512, MN167513, MN167518, MN167519, and MN167520-9). The European haplotypes are detailed in Roe et al., 2017 (13), and the African Americans are detailed in the previously mentioned manuscript by Roe et al. that has been submitted for peer review. The distribution of haplotype structures in the European cohort is 8 cA01~tA1, and 1 each of cA01~tB01, cA01~tB04, cA02~tA03, cA03~tB02, cB01~tB01, cB02~tA01, and cB04~tB03; one individual is homozygous for cA01~tA01 to within a few variants. The distribution of African American haplotypes is 5 cA01~tA01, 3 cB01~tA01, 3 cB03~tA01, 2 cB01~tB01, 1 cA01~tA02, 1 cA01~tB01, and 1 cB02~tA01. Further details are provided in **Supplementary Table 1**.

## Whole Genome Sequencing

Theoretically, if KIR reads could be removed from WGS, the workflow should be able to assemble haplotypes the same as from targeted sequences. To test this hypothesis, whole genome CCS reads were obtained for an Ashkenazim individual (isolate NA24385) from the Genome In a Bottle (GIAB) consortium, as described in Wenger et al. (34). KIR ground truth was unknown previously. KIR reads were separated from WGS as described above and from there the workflow proceeded as usual from the error correcting step (**Figure 2B**).

# RESULTS

Assemblies were evaluated with ground truth in 16 individuals (32 haplotypes) comprising 11 distinct haplotype structures. They were compared with the reference sequences shown in **Figure 3**, which depicts the 11 structures as connections between the same genes in different haplotypes and shows how the structures represent expansion and contraction of the A and B haplotype categories across the *KIR3DP1-KIR2DL4* hotspot. **Table 1** shows the results of the assembly compared with the reference sequences. For the 8 Europeans on average, the full set of 200 candidate probes provided 98% coverage, with 99.98% concordance, and it took 1.1 haplotigs to cover 75% of the reference (LG75). When a set of capture probes was reduced to a select 15 and evaluated on both cohorts, the European coverage lowered to 93%, with the same concordance rate (99.98%), and 1.3 LG75. The results for African Americans were very similar: 92% coverage, 99.98% concordance, and 1.6 LG75. When a select 3 more probes were added for a total of 18 capture probes, the assemblies for the 8 African Americans improved to 97% coverage, with 99.97% concordance, and 1.8 LG75. Most of the missing coverage occurred at the 3' end of the haplotypes: in certain *KIR3DL2* alleles and some sequences extending 3' past *KIR3DL2*. **Figure 4** shows the alignment of the haplotigs relative to the reference haplotype from the same individual MN167513 (cA01~tB01, **Figure 3**) in the 18-probe experiment. It shows a small <2 kb gap in the assembly in *KIR2DL3*. Otherwise, the haplotigs provide complete and overlapping coverage across the reference haplotype sequence. When all the haplotigs are aligned the reference, the statistics report that it takes 2 haplotigs to assembly 75% of the haplotype and that total coverage is 98.4% with 99.98% concordance with the reference. The haplotigs are colored by base (ACGT) and indicate the haplotigs are concordant. Every gene is spanned by at least one haplotig, and all loci are phased with overlapping haplotigs, which the exception of the gap in *KIR2DL3*. **Supplementary Figure 1** contains the assemblies for all individuals in all three sets of experiments, along with NanoPlot, Qualimap, and Quast reports. The reports contain different visualizations and collections of statistics like number and percentage of mapped/unmapped reads, min/max/mean read lengths, ACGT content, coverage, mapping quality, mismatch rates, and indel rates.

The optimized 18 capture probe provided results very similar to the full candidate set of 200. Since this is the most efficient method, this probe coverage is further explained below. The 18 probes covered the haplotypes to an average distance of 2,398 bases. **Figure 5** shows how the probes are distributed across a typical 19 kb region. The image shows an alignment displayed in Integrative Genomics Viewer (IGV) of the set of 18 probes to cB01~tB01 (GenBank reference KP420442). The top of the image shows it is zoomed into 49 kb of the haplotype (~50–100kb). In the middle track, the vertical ticks with the red numbers above indicate the alignment locations of the probes, with the red number being the label of the probe. In the bottom two tracks, the horizontal blue lines indicate the locations of exons (second from the bottom) and repetitive elements (bottom). The probe locations avoid the blue variable (exons) and repetitive (Alus, LINEs, etc.) regions but achieve complete coverage to a resolution of less than 5 kb across the 49 kb. Only 7 distinct probes align to this region. From left to right, the probe sequence 4-3-12-10-2-7-13 occurs three times, except probe 10 does not align in the middle group. This alignment demonstrates how homology can be used to capture continuous KIR DNA over long distances with few probes without capturing off-KIR DNA. The set of 18 probes are included in **Supplementary Table 2**.

The CCS reads in the 18-probe experiment provided an average of 47x coverage of the haplotypes, except for a small gap in all alleles of *KIR2DL2* and *KIR2DL3*, and a few alleles in other genes such as *KIR2DS2*. The gaps were on average ~100 bases long, lead to gaps in the assembly <2 kb, and were most likely introduced during PCR amplification of a repeat-rich region. See the reports in **Supplementary Figure 1** for more information.

Using the 18 sequences as virtual probes to capture KIR reads from WGS, the KIR assembled into a paternal cA01~tB01 (KIR3DL3*00101~KIR2DL3*00101~KIR2DP1*NEW~KIR2DL1*00302~KIR3DP1*0030202~KIR2DL4*0050101~KIR3DS1*01301~KIR2DL5A*00101~KIR2DS5*00201~KIR2DS1*00201~KIR3DL2*00701) haplotype and a maternal cB05~tB01 (KIR3DL3*00301~KIR2DS2*005~KIR2DP1*NEW~KIR2DL1*0040105~KIR3DP1*0030202~KIR2DL4*00501~KIR3DS1*01301~KIR2DL5*00101~KIR2DS5*00201~KIR2DS1*00201~KIR3DL2*NEW); the assembly and its annotations are included in **Supplementary Figure 2**. The distal/telomeric halves (*KIR3DP1-KIR2DS1*) are mostly homozygous, with approximately a dozen variants between them. It is possible this region is really deleted in maternal haplotype and should be classified as cB05~tB02 (*KIR3DL3~KIR2DS2/KIR2DS3~KIR2DP1~KIR2DL1/KIR2DS2~KIR3DL2*). Either way, this is the first reference haplotype with a cB05 with tB01 or tB02 in the same

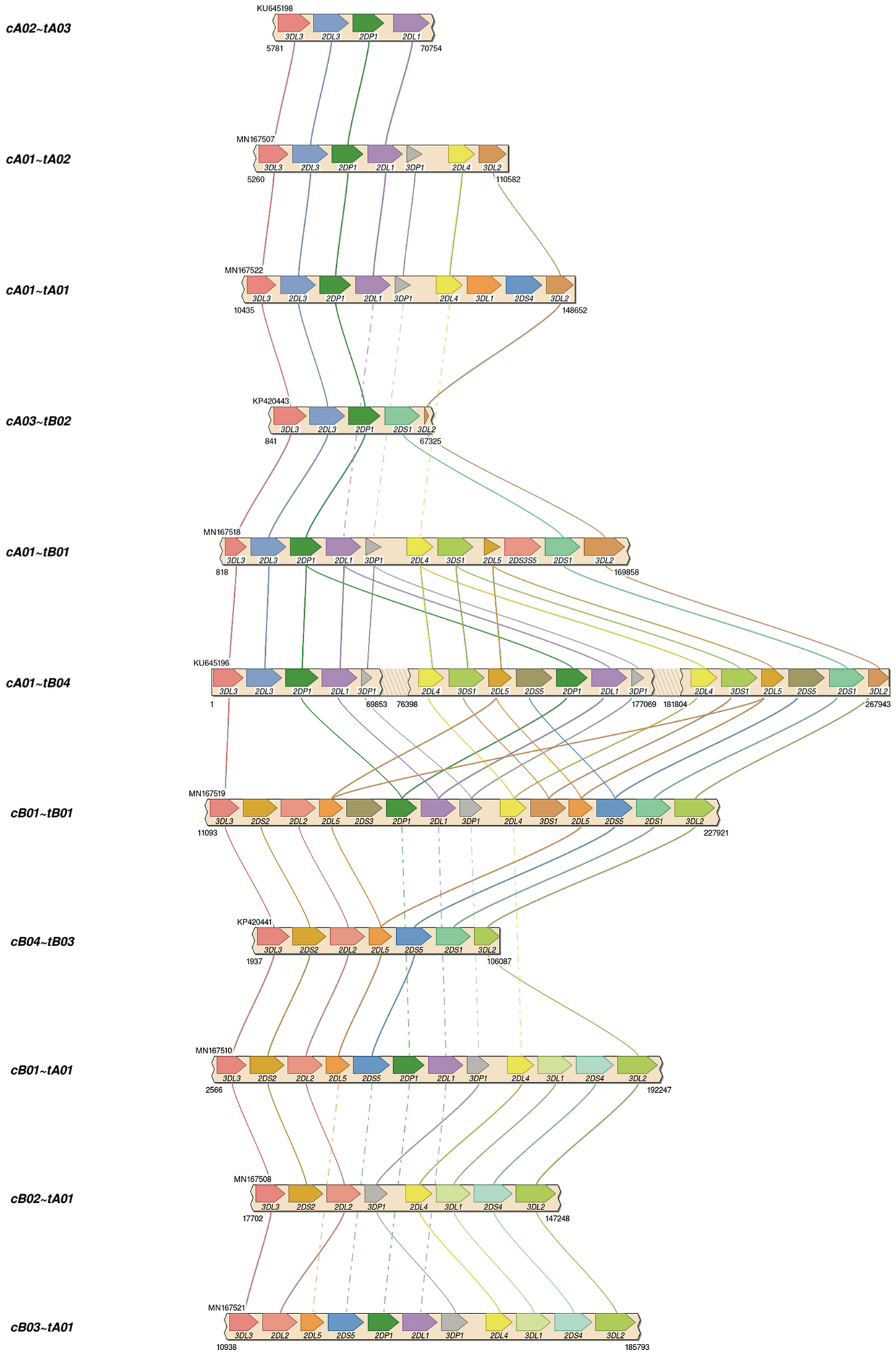

cA02~tA03
KU645198
3DL3 2DL3 2DP1 2DL1
5781 70754
cA01~tA02
MN167507
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DL2
5260 110582
cA01~tA01
MN167522
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DL1 2DS4 3DL2
10435 148652
cA03~tB02
KP420443
3DL3 2DL3 2DP1 2DS1 3DL2
841 67325
cA01~tB01
MN167518
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DS1 2DL5 2DS3S5 2DS1 3DL2
818 169858
cA01~tB04
KU645196
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DS1 2DL5 2DS5 2DP1 2DL1 3DP1 2DL4 3DS1 2DL5 2DS5 2DS1 3DL2
1 69853 76398 177069 181804 267943
cB01~tB01
MN167519
3DL3 2DS2 2DL2 2DL5 2DS3 2DP1 2DL1 3DP1 2DL4 3DS1 2DL5 2DS5 2DS1 3DL2
11093 227921
cB04~tB03
KP420441
3DL3 2DS2 2DL2 2DL5 2DS5 2DS1 3DL2
1937 106087
cB01~tA01
MN167510
3DL3 2DS2 2DL2 2DL5 2DS5 2DP1 2DL1 3DP1 2DL4 3DL1 2DS4 3DL2
2566 192247
cB02~tA01
MN167508
3DL3 2DS2 2DL2 3DP1 2DL4 3DL1 2DS4 3DL2
17702 147248
cB03~tA01
MN167521
3DL3 2DL2 2DL5 2DS5 2DP1 2DL1 3DP1 2DL4 3DL1 2DS4 3DL2
10938 185793

**FIGURE 3** | Reference haplotype structures in the two validation cohorts. Each haplotype represents one of the structures previously established *via* fosmid library preparation and long-read sequencing. The unofficial name of the haplotype is on the left. Lines connect genes with the same name in different structures. Solid lines connect the same gene in neighboring structures. Dashed lines connect the same gene in non-neighboring structures (i.e., the line goes through one or more neighboring haplotypes). cA02~tA03: 1 EUR. cA01~tA02: 1 AFA. cA01~tA01: 5 AFA, 9 EUR. cA03~tB02: 1 EUR. cA01~tB01: 1 AFA, 1 EUR. cA01~tB04: 1 EUR. cB01~cB01: 2 AFA, 1 EUR. cB04~tB03: 1 EUR. cB01~tA01: 3 AFA. cB02~tA01: 1 EUR. cB03~tA01: 1 AFA.

**TABLE 1** | Assembly statistics.

| pop. | probes # | LG75 | coverage % | concordance % |
|---|---|---|---|---|
| EUR | 200 | 1.1 | 98% | 99.98% |
| EUR | 15 | 1.3 | 93% | 99.98% |
| AFA | 15 | 1.6 | 92% | 99.98% |
| **AFA** | **18** | **1.8** | **97%** | **99.97%** |

*For each of experiments in the rows, shown is the population, number of capture probes, the number of haplotigs spanning 75% of the haplotype (LG75), and the coverage of and concordance to the reference. The most efficient experiment is bolded.*

haplotype; cB05 has *KIR2DS2*/*KIR2DS3* fusion (named KIR2DS2*005 in IPD-KIR). The preprinted algorithm to genotype KIR from WGS confirms cB05~tB01 with cA01~tB01 (or their deleted forms) as the most-likely pair of structural haplotypes. The recently published PacBio full-genome assembly of this individual (34) assembled the cA01~tB01 in paternal haplotig SRHB01000968.1, but the maternal haplotigs do not contain any KIR haplotypes.

The code to assemble and annotate KIR haplotypes from CCS reads, including an example, is located at https://github.com/droeatumn/kass. The "main" workflow performs the assembly. The "annotate" workflow labels the genes, exons, and introns in GenBank's.tbl format. The "align" workflow aligns the haplotigs to a reference and produces reports with which to evaluate the assembly or raw data. The code is supported by a Docker container at https://hub.docker.com/repository/docker/droeatumn/kass, for convenient execution. The minimum recommended hardware for targeted sequencing is 30G main memory and 8 CPU cores. More of each is helpful, especially with WGS. On an Ubuntu 18.04 Linux server with 40 core (Intel Xeon CPU E5-2470 v2 @ 2.40GHz) and 132G main memory, a single targeted assembly (~70M fastq.gz) averaged 66 minutes and the WGS (~70G fastq.gz) was 69 minutes. On MacOS 10.15.5 with 4 core (2.7 GHz Quad-Core Intel Core i7) and 16G main memory, a single targeted assembly averaged 125 minutes. Average times are reduced when assemblies are run in parallel.

## DISCUSSION

These experiments in individuals from diverse populations demonstrate that KIR haplotypes can be efficiently enriched and ultimately assembled using an efficient number of capture probes. The workflow successfully reconstructed both haplotypes from targeted sequencing in 16 individuals and from WGS in 1 individual. Although recent advances in Single-Molecule Real-Time sequencing by PacBio have improved quality and extended its applicability to WGS and highly repetitive regions, these advancements are not sufficient to accurately assemble KIR haplotypes. In our evaluations of self-reported results from other published and pre-published assemblers, we could not find another assembler that correctly assembled diploid KIR haplotypes from PacBio reads alone. The sequences need to be error corrected, separated from off-KIR, and sometimes the reads need to be separated and assembled on a per-gene basis depending on depth and genotypic variation, as more binning helps overcome the challenges of higher multiplexing. We have confirmed successful assembly with multiplexing up to eight individuals, which, we estimate, should lead to costs that rival full-exon short-read sequencing, and an order-of-magnitude

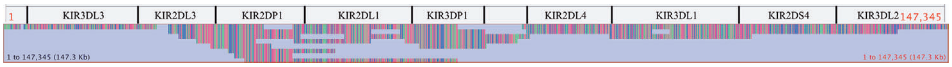


**FIGURE 4** | Alignment of assembled haplotigs with reference haplotype sequence MN167513 (cA01~tA01), whose length is 147,345. The gene features are annotated across the top. The haplotigs are stacked below and colored by nucleotide.

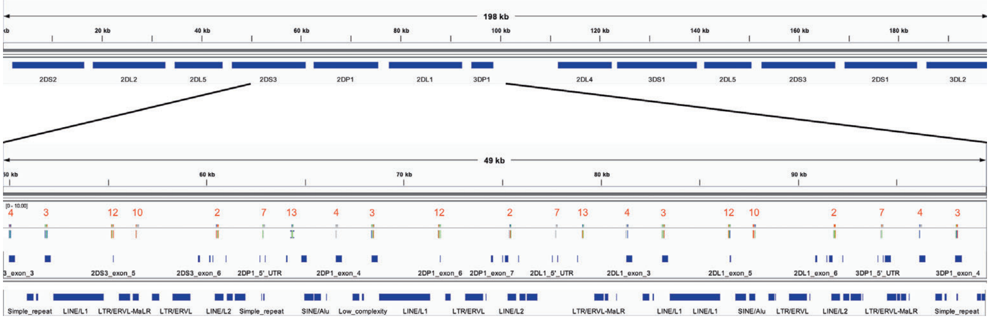


**FIGURE 5** | IGV depiction of the alignment of the 18 set capture probes across a 49 kb region of KP420440 (cB01~tB01). The locations of the probes are displayed by the vertical ticks with the red labels above them in the middle track. The locations of exons and repeat elements (horizontal blue bars) are on the bottom two tracks. Seven distinct probes align in this window. The probe pattern 4-3-12-10-2-7-13 repeats three times from left to right, except probe 10 does not align in the middle group.

more efficient than fosmid-based library preparation due to a more high-throughput library preparation workflow. We report efficiency details as a minimum, because the purpose of these experiments were to demonstrate the capability, not maximum efficiency. Our results suggested that lower bounds efficiency for off/on ratio for KIR capture range between 1.3 to 2.4 (**Supplementary Table 3**).

In addition to assembly, our software is the first annotation system for KIR haplotypes. The annotation algorithm leverages the information of the pattern of capture probes across the haplotype sequence to define loci and their exon/intron locations. Another system annotates gene alleles (35), but only if the input allele sequences start at the same location as their PCR primers.

There are 50 KIR reference haplotypes in the human genome reference GRCh38.p13 (36). The 16 African American haplotypes will be added in the next release, bringing the total to 66. This is 3 times more than the entire chromosome 3, which has the next highest total number of alternative haplotypes. Almost half of these KIR haplotype sequences were characterized in two workflows whose only common step was PacBio sequencing. Since both workflows agree over 99%, we can be confident the cohorts have been characterized correctly and the two approaches validated each other.

A 2016 manuscript (37) describes PING, which is software to interpret KIR from short (<= 300 bp) reads. It uses probes to capture 800 bp DNA fragments. Although the total number of probes required for KIR capture was unspecified, the total number for KIR and HLA was 10,456. PING's highest resolution results are obtained by aligning the short reads to full-gene references, calling the SNP variants, and then calling the two most likely reference alleles given the SNP genotypes. Although a great improvement over other current technologies at the time, the PING method does not phase/link variants within a gene or the haplotype as our long-read sequencing and capture method allows. Further, PING uses more than an order of magnitude more capture probes, which can be expensive to capture shorter fragments. It produces probabilistically phased lower-resolution predictions compared with our long-read assembly, which produces linked multi-gene and haplotype sequences without references. Therefore, we feel our method appreciably adds to the ability to properly analyze KIR regions by leveraging long read technologies for increased resolution and phasing compared with the previous NGS approach and also by leveraging high-throughput library preparation for reduced cost compared with the previous haplotyping approach.

In addition to demonstrating an efficient targeted haplotyping strategy, to the best of our knowledge, this the first report of KIR full diploid haplotype assembly from HiFi WGS alone. Our approach was able to assemble both haplotypes from WGS whereas the previously reported whole-genome assembly could not, underlying the necessity of a KIR-specific assembler. Both regions comprising the haplotype it missed (cB05~tB01 or a deleted form) are not in the primary human genome reference, and the two have not been reported together previously. Perhaps this lack of representation in the reference contributed to the missing assembly. Other possibilities include the lack of binning/separation of KIR reads from the rest of the genome before assembling, or differences in the tools used in the workflow. Regardless, this experiment demonstrates the value added by the bioinformatics algorithms, in addition to the targeted capture and assembly.

The suspected amplification problem causing the small gap in KIR2DL2L3 occurs in a ~100 base region of poly-ATs, with L1s (and ALUs) on either side. It appears most or all PCR methods have a problem with this region, as almost every *KIR2DL2* and *KIR2DL3* reference allele has a different poly-AT sequence for this region (**Supplementary Figure 3**); these reference alleles were sequenced on various platforms but generally (if not fully) amplified with PCR. The PCR-less WGS from GIAB have no gaps, which suggests the source of the gap is PCR amplification and demonstrates that the assembler can correctly assemble this region along with the rest of the haplotypes when using non-PCR library preparation methods. Since *KIR2DL2* or *KIR2DL3* occur in most haplotypes and occur ~10–20% from the proximal end, their short gap limits LG75 to its reported value of 1.8 in the AFA cohort. In cases where the gap is not an issue, such as WGS, LG75 will probably be 1, and LG100 will probably be a better metric.

The power of this method to assemble repetitive KIR regions without incorporating false non-KIR genomic signals may lie in the strongest recombination hotspot, the 10 kb intervening region between *KIR3DP1* and *KIR2DL4*. Conventionally, KIR haplotype names (e.g., "cA01~tA01") have been described as two halves ("c" and "t") separated by a recombination hotspot ("~"). The rate of recombination between the two halves is so frequent that any two may be found with each other, despite the relative evolutionary youth of the region. This hotspot stretches over 9 kb between *KIR3DP1* and *KIR2DL4*. Any two alleles of this region are over 99% identical and consist of 13% Alus (SINEs) and 58% LINE1 repeat elements. **Figure 6** shows an alignment of the region to itself. The top part of the figure displays the location of the *KIR3DP1-KIR2DL4* intergenic region in the context of the cA01~tA01 primary human genome reference. The bottom half zooms into the intergenic region and shows a dot plot of the alignment. The lines on the dot plot indicate stretches of the haplotype that align with itself, either in the same location or a different location. The red lines indicate matching in the same orientation as the overall haplotype, and the blue indicate matching in the reverse complement. The red and blue horizontal bars at the bottom of the figure detail the location of repetitive elements. The red from 0–2,000 simply shows that this region aligns with itself. There is a stretch of 3 kb from ~5,200–8,200 that reverse complement matches ~400–3,400. The yellow boxes highlight Alu repeats: AluSx3 from ~800–1,000 is matched in the reverse complement by AluSx1 (~6,400–6,600) and AluSx4 (~7,100–7,400) and the same orientation by AluSq2 (~8,200–8,400). All elements are surrounded by very similar L1 elements for at least 1,000 bp on both sides. This stretch of 3,000+ bases of reverse complement repeats provides fertile ground for homologous recombination between the two halves of KIR haplotypes. This is the most difficult region to phase and

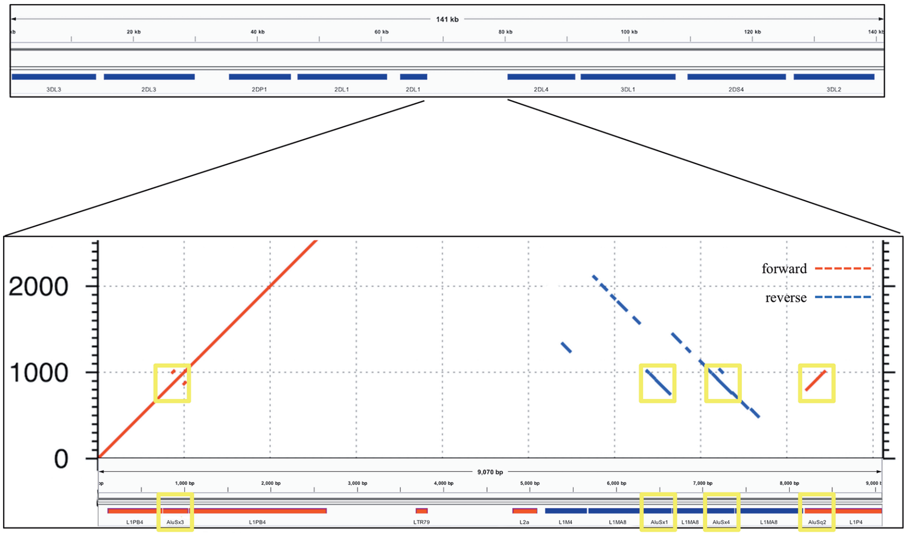


**FIGURE 6** | Recombination hotspot. The top shows the 9 kb haplotype context of the repetitive region of the *KIR3DP1-KIR2DL4* intergene region, whose self-alignment is shown in the bottom half. Red lines indicate alignment of the haplotype with itself in the same orientation, and blue lines indicate reverse complement orientation. The location of the repetitive elements is at the bottom, with red and blue again indicating orientation. The yellow boxes highlight three AluSx and one AluSq elements that align with each other, two in each direction.

the results demonstrate that the combination of variants and read length is generally high enough to phase full haplotypes with diploid reads. Although this region is an extreme example, the other recombination sites, which are usually internal to genes and result in gene fusions, homologously recombine *via* the same elements (16, 38–41), although only the *KIR2DL4-KIR3DP1* intergenic region contains as many elements in both directions (10). Our assay is the only high-throughput method that allows analysis of all of these regions.

Future efforts include expanding testing to other populations, resolving the KIR2DL2L3 gap if possible, expanding capture for some *KIR3DL2* alleles, expanding the assembly and annotation to bordering genes in the leukocyte receptor complex, optimizing multiplexing, and incorporate scaffolding into the workflow. Although the diverse AFA and EUR cohorts demonstrate proof of concept and expand our human genome references, it is important to develop reference sets for all populations. Expanding the capture would help ensure that *KIR3DL3* and *KIR3DL2* are sufficiently captured, help define any deleted haplotypes that may include these two genes and capture potentially relevant regulatory signals. Currently, all fully sequenced haplotypes contain some portion of these two bordering genes. “Scaffolding” is the term for combining haplotigs into one final haplotype sequence. Although the experiments revealed most haplotypes are covered in only two haplotigs and the haplotigs are simply subsequences of other haplotigs, it might help downstream analysis to include a rigorous scaffolding step; the KPI software may help select the appropriate references, as it can predict the pair of haplotype structures from the raw sequences.

Our approach leverages sequence similarity across multiple loci that were created by duplication followed by variation. Since this this is a true for many gene families, our approach should be more generally applicable to other regions that have a mix of homologous and variable/repetitive regions relative fragment length and capture characteristics.

The application of KIR genetics in medical research such as immunity, reproduction, and transplantation is encouraging, but limited by the technical difficulties for high-resolution interpretations at large scale and low cost. Here, a KIR haplotyping workflow was presented that can provide full-sequence haplotypes at approximately the same cost as full exon or full gene. For the first time, it allows high-resolution KIR haplotypes in population-sized cohorts, as opposed to lower-resolution genotypes. The analysis pipeline uses domain knowledge to assemble reads generated *via* well-established sequencing techniques that is accurate enough for personalized precision medicine and scalable to populations. To this point, most KIR association studies focus on variation at only one locus or one functional class to associate, while keeping the rest of the haplotypes static. Future full-haplotype studies will help KIR researchers better study gene combinations, regulatory regions, recombination hotspots, self-regulation, and non-binding factors that influence disease phenotypes. This increased ability will provide completed sets of population-specific reference haplotypes which will, among other things, enhance imputation power of lower resolution data. It allows for new comparisons that will provide insight into evolution and make this region the best annotated in the human genome, despite its complexity. Lastly, this novel approach will provide the capability to discover genetic associations in medically relevant areas such as infections, transplantation, cancer susceptibility, autoimmune diseases, reproductive conditions, and immunotherapy. The open and free software is available at https://github.com/droeatumn/kass and supported by a environment at https://hub.docker.com/repository/docker/droeatumn/kass.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material** or in the input directory at https://github.com/droeatumn/kass.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by National Marrow Donor Program Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

DR, JW, RK, and MM designed the experiments. DR, JW, KI, JB, ND, and CL performed the experiments. All authors contributed to the article and approved the submitted version.

## FUNDING

Supported by a grant from the Department of the Navy, Office of Naval Research (N00014-19-1-2888).

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv with the same title and authors (42). Thanks to Cynthia Vierra-Green from CIBMTR, Julia Udell from BICB UMN, and Richard Hall from PacBio for advice and review.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.582927/full#supplementary-material

**SUPPLEMENTARY FIGURE 1 |** AFA and EUR haplotigs. Data Sheets 2-5 are zip files containing the assembled haplotigs for all AFA and EUR assemblies. Also included are Qualimap, NanoPack, and QUAST reports.

**SUPPLEMENTARY FIGURE 2 |** WGS haplotigs. Data Sheet 1 is a zipped file containing the assembled haplotigs for maternal and paternal haplotypes for WGS assembly from GIAB individual NA24385.

**SUPPLEMENTARY FIGURE 3 |** Image 3.tif contains a multiple sequence alignment of reference *KIR2DL2* alleles in IPD-KIR. The variation between columns 6,706 and 6,744 demonstrates extensive reported variation in a poly-AT region.

**SUPPLEMENTARY TABLE 1 |** Cohort details. In Table 1, the first column contains the individual ID used for this study. The second column contains the GenBank accession number of the reference haplotype. Accessions that start with K are from the European cohort, and accessions that start with M are form the African American cohort. The third column contains the informal haplotype name.

**SUPPLEMENTARY TABLE 2 |** Capture probes. Data sheet 6 contains the 18 capture probe sequences in zipped fasta format.

**SUPPLEMENTARY TABLE 3 |** The spreadsheet Table 2 contains statistics calculating the amount and ratios of on and off KIR sequences in the experiments that differ by cohort and number of probes.

## REFERENCES

1. Wroblewski EE, Parham P, Guethlein LA. Two to Tango: Co-evolution of Hominid Natural Killer Cell Receptors and MHC. *Front Immunol* (2019) 10:177. doi: 10.3389/fimmu.2019.00177
2. Rajalingam R. Human diversity of killer cell immunoglobulin-like receptors and disease. *Korean J Hematol* (2011) 46:216. doi: 10.5045/kjh.2011.46.4.216
3. Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, et al. Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nat Genet* (2007) 39:733–40. doi: 10.1038/ng2035
4. Davis C, Rizzieri D. Immunotherapeutic Applications of NK Cells. *Pharmaceuticals* (2015) 8:250–6. doi: 10.3390/ph8020250
5. Foley B, Felices M, Cichocki F, Cooley S, Verneris MR, Miller JS. The biology of NK cells and their receptors affects clinical outcomes after hematopoietic cell transplantation (HCT). *Immunol Rev* (2014) 258:45–63. doi: 10.1111/imr.12157
6. Benson DM, Caligiuri MA. Killer Immunoglobulin-like Receptors and Tumor Immunity. *Cancer Immunol Res* (2014) 2:99–104. doi: 10.1158/2326-6066.CIR-13-0219
7. Moffett A, Colucci F. Co-evolution of NK receptors and HLA ligands in humans is driven by reproduction. *Immunol Rev* (2015) 267:283–97. doi: 10.1111/imr.12323
8. Leaton LA, Shortt J, Kichula KM, Tao S, Nemat-Gorgani N, Mentzer AJ, et al. Conservation, Extensive Heterozygosity, and Convergence of Signaling Potential All Indicate a Critical Role for KIR3DL3 in Higher Primates. *Front Immunol* (2019) 10:24. doi: 10.3389/fimmu.2019.00024
9. IPD-KIR. https://www.ebi.ac.uk/ipd/kir/.
10. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci* (2000) 97:4778–83. doi: 10.1073/pnas.080588597
11. Pyo C-W, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different Patterns of Evolution in the Centromeric and Telomeric Regions of Group A and B Haplotypes of the Human Killer Cell Ig-Like Receptor Locus. *PloS One* (2010) 5:e15115. doi: 10.1371/journal.pone.0015115
12. Pyo C-W, Wang R, Vu Q, Cereb N, Yang SY, Duh F-M, et al. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics* (2013) 14:89. doi: 10.1186/1471-2164-14-89
13. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun* (2017) 18(3):127–34. doi: 10.1038/gene.2017.10
14. Schwartz JC, Gibson MS, Heimeier D, Koren S, Phillippy AM, Bickhart DM, et al. The evolution of the natural killer complex; a comparison between mammals using new high-quality genome assemblies and targeted annotation. *Immunogenetics* (2017) 69:255–69. doi: 10.1007/s00251-017-0973-y
15. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Res* (2012) 22:1845–54. doi: 10.1101/gr.137976.112

16. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet* (2010) 19:737–51. doi: 10.1093/hmg/ddp538
17. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, et al. Human Diversity in Killer Cell Inhibitory Receptor Genes. *Immunity* (1997) 7:753–63. doi: 10.1016/S1074-7613(00)80394-5
18. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* (2014) 43(1):D423–31. doi: 10.1093/nar/gku1161
19. RepeatMasker. http://www.repeatmasker.org/.
20. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* (2015) 6:11. doi: 10.1186/s13100-015-0041-9
21. Multiplex Genomic DNA Target Capture Using IDT xGen® Lockdown® Probes. Available at: https://www.pacb.com/wp-content/uploads/Procedure-Checklist-%E2%80%93-Multiplex-Genomic-DNA-Target-Capture-Using-IDT-xGen-Lockdown-Probes.pdf (Accessed August 31, 2020).
22. Salmela L, Walve R, Rivals E, Ukkonen E. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* (2016) 799–806. doi: 10.1093/bioinformatics/btw321
23. Roe D, Kuang R. Accurate and Efficient KIR Gene and Haplotype Inference from Genome Sequencing Reads with Novel K-mer Signatures. *bioRxiv* (2019). doi: 10.1101/541938
24. BBTools. . http://sourceforge.net/projects/bbmap/.
25. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* (2017) 27:722–36. doi: 10.1101/gr.215087.116
26. Lafita A, Bliven S, Prlić A, Guzenko D, Rose PW, Bradley A, et al. BioJava 5: A community driven open-source bioinformatics library. *PloS Comput Biol* (2019) 15:e1006791. doi: 10.1371/journal.pcbi.1006791
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* (2018) 34:3094–100. doi: 10.1093/bioinformatics/bty191
28. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* (2015) 292–4. doi: 10.1093/bioinformatics/btv566
29. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* (2018) 34:2666–9. doi: 10.1093/bioinformatics/bty149
30. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* (2013) 29:1072–5. doi: 10.1093/bioinformatics/btt086
31. Veltri D, Wight MM, Crouch JA. SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. *Nucleic Acids Res* (2016) 44:W41–5. doi: 10.1093/nar/gkw330
32. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. *Briefings Bioinf* (2013) 14:193–202. doi: 10.1093/bib/bbs012
33. Roe D, Vierra-Green C, Pyo C-W, Geraghty DE, Spellman S, Kuang R, et al. A Detailed View of KIR Haplotype Structures and Gene Families as Provided by a New Motif-based Multiple Sequence Alignment. *bioRxiv* (2020). doi: 10.1101/2020.08.07.242305
34. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* (2019) 37:1155–62. doi: 10.1038/s41587-019-0217-9
35. Surendranath V, Albrecht V, Hayhurst JD, Schöne B, Robinson J, Marsh SGE, et al. TypeLoader: A fast and efficient automated workflow for the annotation and submission of novel full-length HLA alleles: SURENDRANATH et al. *HLA* (2017) 90:25–31. doi: 10.1111/tan.13055
36. KIR haplotypes in the human genome reference. *Genome Reference Consortium*. (2020). Available at: https://www.ncbi.nlm.nih.gov/grc/human?filters=chr:19+gene:KIR#current-regions.
37. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *Am J Hum Genet* (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023
38. Martin AM, Freitas EM, Witt CS, Christiansen FT. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* (2000) 51:268–80. doi: 10.1007/s002510050620
39. Martin MP, Bashirova A, Traherne J, Trowsdale J, Carrington M. Cutting Edge: Expansion of the *KIR* Locus by Unequal Crossing Over. *J Immunol* (2003) 171:2192–5. doi: 10.4049/jimmunol.171.5.2192
40. Sambrook JG. Single haplotype analysis demonstrates rapid evolution of the killer immunoglobulin-like receptor (KIR) loci in primates. *Genome Res* (2005) 15:25–35. doi: 10.1101/gr.2381205
41. Guethlein LA, Older Aguilar AM, Abi-Rached L, Parham P. Evolution of Killer Cell Ig-Like Receptor ( *KIR* ) Genes: Definition of an Orangutan *KIR* Haplotype Reveals Expansion of Lineage III KIR Associated with the Emergence of MHC-C. *J Immunol* (2007) 179:491–504. doi: 10.4049/jimmunol.179.1.491
42. Roe D, Williams J, Ivery K, Brouckaert J, Downey N, Locklear C, et al. Efficient Sequencing, Assembly, and Annotation of Human KIR Haplotypes. *bioRxiv* (2020). doi: 10.1101/2020.07.12.199570

**Conflict of Interest:** JW, KI and JB were employed by Laboratory Corporation of America Holdings. ND and CL were employed by Integrated DNA Technologies Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Immunology

ORIGINAL RESEARCH
published: 06 November 2020
doi: 10.3389/fimmu.2020.584950

# How Ancestry Influences the Chances of Finding Unrelated Donors: An Investigation in Admixed Brazilians

Kelly Nunes[1*], Vitor R. C. Aguiar[1], Márcio Silva[2], Alexandre C. Sena[2], Danielli C. M. de Oliveira[3], Carla L. Dinardo[4], Fernanda S. G. Kehdy[5], Eduardo Tarazona-Santos[6], Vanderson G. Rocha[4,7], Anna Barbara F. Carneiro-Proietti[8], Paula Loureiro[8,9], Miriam V. Flor-Park[10], Claudia Maximo[11], Shannon Kelly[12,13], Brian Custer[12,14], Bruce S. Weir[15], Ester C. Sabino[16], Luís Cristóvão Porto[17] and Diogo Meyer[1*]

*[1] Laboratory of Evolutionary Genetics, Institute of Biosciences, University of São Paulo, São Paulo, Brazil, [2] Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil, [3] Registro Nacional de Doadores Voluntários de Medula Óssea—REDOME, Instituto Nacional do Câncer, Ministério da Saúde, Rio de Janeiro, Brazil, [4] Fundação Pró Sangue, Hemocentro de São Paulo, São Paulo, Brazil, [5] Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [6] Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [7] Serviço de Hematologia, Hemoterapia e Terapia Celular, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil, [8] Fundação Hemominas, Belo Horizonte, Brazil, [9] Fundação de Hematologia e Hemoterapia de Pernambuco, HEMOPE, Recife, Brazil, [10] Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Instituto da Criança, São Paulo, Brazil, [11] Fundação Hemorio, Rio de Janeiro, Brazil, [12] Epidemiology, Vitalant Research Institute, San Francisco, CA, United States, [13] University of California San Francisco Benioff Children's Hospital Oakland, Oakland, CA, United States, [14] Department of Laboratory Medicine, University of California San Francisco, San Francisco, CA, United States, [15] Department of Biostatistics, University of Washington, Seattle, WA, United States, [16] Instituto de Medicina Tropical, Departamento de Moléstias Infecciosas e Parasitárias da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil, [17] Laboratório de Histocompatibilidade e Criopreservação, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil*

OPEN ACCESS

***Edited by:***
*Martin Maiers,*
*National Marrow Donor Program,*
*United States*

***Reviewed by:***
*Pierre-Antoine Gourraud,*
*Université de Nantes, France*
*Abeer Madbouly,*
*Center for International Blood and Marrow Transplant Research (CIBMTR), United States*

****Correspondence:***
*Diogo Meyer*
*diogo@ib.usp.br*
*Kelly Nunes*
*knunesbio@gmail.com*

***Specialty section:***
*This article was submitted to Alloimmunity and Transplantation, a section of the journal Frontiers in Immunology*

***Received:** 18 July 2020*
***Accepted:** 05 October 2020*
***Published:** 06 November 2020*

***Citation:***
*Nunes K, Aguiar VRC, Silva M, Sena AC, de Oliveira DCM, Dinardo CL, Kehdy FSG, Tarazona-Santos E, Rocha VG, Carneiro-Proietti ABF, Loureiro P, Flor-Park MV, Maximo C, Kelly S, Custer B, Weir BS, Sabino EC, Porto LC and Meyer D (2020) How Ancestry Influences the Chances of Finding Unrelated Donors: An Investigation in Admixed Brazilians. Front. Immunol. 11:584950. doi: 10.3389/fimmu.2020.584950*

A match of HLA loci between patients and donors is critical for successful hematopoietic stem cell transplantation. However, the extreme polymorphism of HLA loci – an outcome of millions of years of natural selection – reduces the chances that two individuals will carry identical combinations of multilocus HLA genotypes. Further, HLA variability is not homogeneously distributed throughout the world: African populations on average have greater variability than non-Africans, reducing the chances that two unrelated African individuals are HLA identical. Here, we explore how self-identification (often equated with "ethnicity" or "race") and genetic ancestry are related to the chances of finding HLA compatible donors in a large sample from Brazil, a highly admixed country. We query REDOME, Brazil's Bone Marrow Registry, and investigate how different criteria for identifying ancestry influence the chances of finding a match. We find that individuals who self-identify as "Black" and "Mixed" on average have lower chances of finding matches than those who self-identify as "White" (up to 57% reduction). We next show that an individual's African genetic ancestry, estimated using molecular markers and quantified as the proportion of an individual's genome that traces its ancestry to Africa, is strongly associated with reduced chances of finding a match (up to 60% reduction). Finally, we document that the strongest reduction in chances of finding a match is associated with having an MHC region of exclusively African ancestry (up to 75% reduction). We apply our findings to a specific condition, for which there is a clinical indication for transplantation: sickle-cell disease. We show that the increased African ancestry in patients with this disease leads to reduced chances of finding a match, when compared to the remainder of the sample, without the condition. Our results underscore the influence of ancestry on chances of finding compatible HLA matches, and indicate that efforts guided to increasing the African component of registries are necessary.

**Keywords: HLA, hematopoietic stem cell transplantation, genetic ancestry, admixture, MHC, Brazil**

## INTRODUCTION

Since the first allogeneic transplant in the late 1950s, there has been significant progress in the technical procedures and success rate of hematopoietic stem cell transplantation (HSCT). Nowadays, HSCT has become the standard treatment for several hematological diseases such as hematologic malignancies (leukemias, Hodgkin's lymphoma), and also a curative treatment for some congenital or acquired disorders of the hematopoietic system (sickle cell anemia, severe aplastic anemia, thalassemia and inborn metabolism errors), as well as a therapeutic option in the treatment of some solid tumors (1, 2).

The success of allogeneic HSCT (allo-HSCT) is in part due to a better understanding of the role of HLA genes in the immune response. Classical HLA genes encode antigen-presenting proteins which are recognized by T-cell receptors. When HLA molecules bind a non-self antigen, cellular and humoral immune responses are triggered. Thus, in allo-HSCT the match between patient and donor HLA is critical, since different HLA alleles can generate distinct antigens detected as non-self during cellular immunological inspection. As a consequence, the patient's immune system sees the donor cells with incompatible HLA as foreign and mounts an immune response, leading to rejection of the HSCT, and/or to graft versus host disease after grafting. Therefore, the gold standard for allo-HSCT is full compatibility between patient and potential donor at the 2 alleles of *HLA-A*, *-B*, *-C*, *-DRB1*, and *-DQB1* (10/10 match) [see review in Tiercy, 2016 (3)]. Nevertheless, in some cases mismatches may be allowed (9/10 match), especially at alleles which do not generate an anti-HLA antibody (4, 5), or when new therapeutic protocols of haploidentical transplant are used, in which the donor and recipient share a common HLA haplotype (6).

The extremely high number of alleles at HLA loci (more than 27,000 classical HLA alleles were described in the Immuno Polymorphism Database IMGT/HLA until July 2020, https://www.ebi.ac.uk/ipd/imgt/hla/stats.html) makes relatives of the patient the first option in searching for a donor. The patient's probability of having the same HLA haplotypes as one of their siblings is 25%, and increases with the number of siblings (43.7% with 2, 57.8% with 3, 68.4% with 4 siblings, etc.) (7). However, according to the USA National Marrow Donor Program (NMDP), only 30% of allo-HSCT donors are chosen from close relatives, with unrelated allo-HSCT donors accounting for about 70% of cases (8). Therefore, public donor registries play a key role in meeting the demand for unrelated donors.

According to the World Marrow Donor Association (WMDA), which includes 53 associated countries, the number of registered unrelated donors exceeds 35 million individuals and 700,000 cord blood units (WMDA, 2019; https://statistics.wmda.info/). Despite this large number, a major concern has been to understand how representative existing registries are of the general population, since the high diversity of HLA loci, their geographical heterogeneity (9, 10), and biases in volunteer recruitment, can result in certain groups within the population having reduced access to donors.

This scenario may be critical in the case of admixed populations, which trace their ancestry to different geographic regions, and has been the focus of several recent studies on how ethnicity influences the probability of finding a donor (8, 11–13). In the USA, an NMDP study found that while 75% of patients who self-identify as White-European found a donor with 7/8 HLA matching, only 16% for African-Americans found a match (8). A Memorial Sloan Kettering Cancer Center study of 7/8 or 8/8 HLA matching found that 78% of patients with Northwestern European ancestry found a donor, while this number fell to 44% for those with South European ancestry and 22% for those with African ancestry (13). Theoretical studies by Bergstrom et al. (11, 14) also found that the probability of not finding a match in the NMDP is highest among African-Americans, when compared to other groups. These findings have been interpreted as an outcome of a combination of factors, including a lower representation of African-Americans in databases, as well as the greater genetic diversity of populations originating in Africa, which decreases the likelihood that two unrelated individuals will share a multi-locus HLA genotype (15).

While the studies discussed above make a convincing case regarding the impact of self-identification upon the chances of finding a donor, they do not address the low correspondence between self-identification and genetic ancestry (16–18). In admixed populations, individuals who self-identify to the same group frequently have extremely variable genetic ancestries. This is expected, since self-identification involves a complex interplay of social and genealogical components (19) and varies among geographic regions in Brazil (20, 21).

Here, we investigate the role of genetic ancestry in determining the chances that an individual will find a match in REDOME (*Registro Nacional de Doadores Voluntários de Medula Óssea*), the Brazilian Marrow Donor Registry. We investigate how three layers of information affect the chances of finding an HLA match: (i) the self-identification; (ii) the genetic ancestry; (iii) the genetic ancestry of the MHC region (**Figure 1**). Brazil harbors the largest number of individuals with African ancestry outside Africa (https://www.slavevoyages.org/) and has one of the most admixed populations in the world. In addition, REDOME is the third largest marrow registry in the world, with more than 5 million registered individuals (as of July 2020). We use samples from two cohorts of Brazilians for which we have information on self-identification, and that were previously genotyped with high density SNP arrays (22, 23), providing a total of 8,037 individuals. For each individual we estimate their genetic ancestry, their ancestry within the MHC, and we impute their HLA genotypes using the SNPs flanking the classical HLA loci. We then query the REDOME to identify how many matches are present for each individual, and use these results to compare how different measures of ancestry influence the chances of finding a match (**Figure 1**).

# MATERIAL AND METHODS

## Datasets

Individuals from two Brazilian cohorts were queried for matches in REDOME, as described below. Although only a subset of individuals in these cohorts are in fact patients for whom there is an indication of transplantation, for the purposes of our analyses we will treat all samples as patients, given that our goal is to establish relationships between ancestry and chances of finding a potential donor among Brazilians with varying ancestries.

The Brazilian National Ethical Committee for Research (CONEP, resolution 15.895 and resolution 1.297.627), local ethics committees at each participating center, the Institutional Review Boards at the University of California, San Francisco and the REDS-III data coordinating center, RTI International, all reviewed and approved the study.

## The Recipient Epidemiology and Donor Evaluation (REDS)-III Brazil Sickle Cell Disease (SCD) Cohort

The REDS-III Brazil SCD cohort was established to study the epidemiology and transfusion outcomes of SCD in Brazil, and includes 2,795 individuals recruited between 2013 and 2015 in 4 reference centers: Fundação Hemominas (Belo Horizonte, Juiz de Fora, and Montes Claros), Fundação Hemope (Recife), Fundação Hemorio (Rio de Janeiro), and Instituto da Criança Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (São Paulo) (**Figure S1**). Samples were previously genotyped with a high density SNP array (Axiom Transfusion Medicine Array, TM Array, Affymetrix, Santa Clara, CA, USA) and are available in the dbGAP (phs001972.v1.p1). After filtering (calling < 97%, Hardy-Weinberg p-value > $10^{-8}$, sample missing data <5%) the dataset had 2,703 individuals and 820,837 SNPs (22).

## EPIGEN Brazil Initiative

The Brazilian EPIGEN Initiative (EPIGEN) is the largest Brazilian resource (n=6,487) for population genomics. EPIGEN includes three cohorts: Salvador (n=1,309), from northeast Brazil

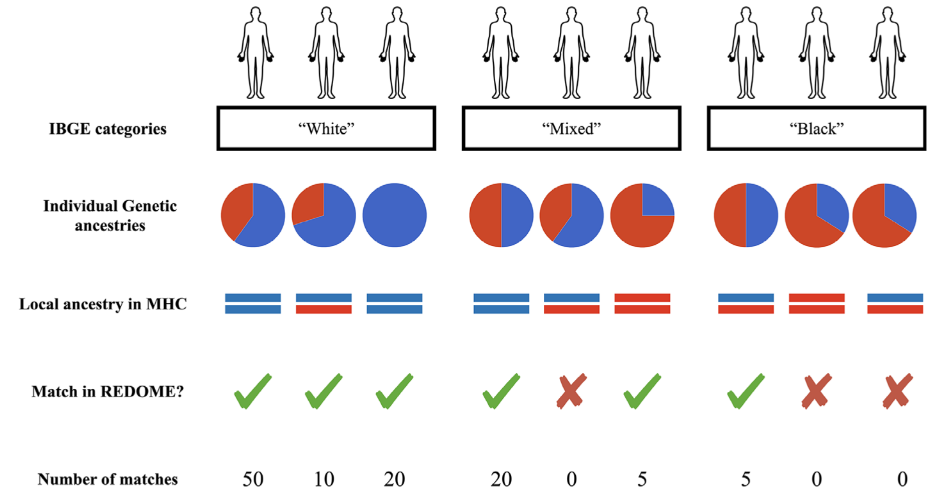


**FIGURE 1** | Schematic representation of the strategy used in this study. Individuals are classified by three criteria: "IBGE categories" are self-identified categories defined by the Brazilian Institute of Geography and Statistics (IBGE), based on phenotypic characteristics such as skin color, and can be "Black", "Mixed", "Indigenous", "White", and "Yellow"; "Genetic Ancestry" is inferred with genetic markers spread throughout the genome, and describes the proportion of an individual's genome that can be assigned to a continent of origin (with red and blue corresponding to African and European proportion ancestries respectively); "MHC ancestry" is the genetic ancestry of the segment of chromosome 6 including the MHC region within an individual (again, with red and blue corresponding to African and European ancestries). For each individual, classified with these criteria, we queried REDOME for the presence of a matching genotype and recorded the number of matches, when they were found.

(24); Bambuí (n=1,442), from the southeast (25); and Pelotas (n=3,736), from south Brazil (26) (**Figure S1**). The samples were previously genotyped with the HumanOmin 2.5 (Illumina, San Diego, CA, USA) SNP array (23) and are available in the European Nucleotide Archive (PRJEB26388 (ERP108374)), under EPIGEN Committee Controlled Access mode. We selected samples for which both genotype data and self-identification were available, resulting in a total of 5,334 samples (Salvador, n = 918; Bambuí, n = 765; and Pelotas, n = 3,651).

## HLA Imputation

For admixed and non-European populations, Attribute Bagging is the most accurate approach for imputing HLA alleles from SNP data (27–29). Here we use the HLA Genotype Imputation with Attribute Bagging (HIBAG) R package (30). HIBAG uses a reference panel with data on both HLA alleles and SNPs in the MHC region to infer HLA alleles for samples with only SNP data.

We built a multi-ancestry imputation model for *HLA-A, -B, -C, -DQB1*, and *-DRB1* with two-field level of resolution using the 1000 Genomes phase III (2,504 individuals) as a reference panel (31, 32). We selected SNPs within the MHC region which are present in both the TM array (10,711 SNPs) and HumanOmin2.5 (9,187 SNPs). For each SNP array, HIBAG models were built using SNPs within 500 kb flanking each HLA locus, and 100 bootstrap samples as classifiers. Out-of-bag estimated accuracies for each model are reported in **Table S1**, and models are available upon request. HLA imputation was performed separately for REDS-III and EPIGEN, and the posterior probability estimated by the model was used as a confidence score, allowing inference of the predicted accuracy of the imputation for each HLA genotype. We used the empirical cumulative distribution function (ECDF) to compare the posterior probability distribution of the imputed HLA genotypes among IBGE categories and genetic ancestries (**Tables S2** and **S3** and **Figures S6** and **S7**).

Since Native American populations are not well represented in the reference panel, HLA imputation for individuals of this ancestry is uncertain, so we chose to focus exclusively on the effects of African and European ancestry in subsequent analyses. Understanding of how Native American ancestry impacts the chances of finding a donor in REDOME will require direct typing of HLA alleles.

## Finding Matches in REDOME

The Brazilian Bone Marrow Donor Registry (REDOME) was established in 1993, funded by the Brazilian Ministry of Health. As of April 30, 2019, REDOME had 4,869,224 registered volunteers. All individuals are genotyped at *HLA-A, -B* and *-DRB1*, and a subset also at *HLA-C* (n=125,248) and *-DQB1* (n=123,298), with HLA resolution ranging between low (e.g., serological assays) to medium/high (e.g., SBT-PCR, Next Generation Sequencing) (**Table S4**).

For each individual, we queried REDOME for potential donors at both low and medium resolutions (one and two HLA allele fields, respectively). At low-resolution an allele such as A*34:02 is compatible with those from donors carrying any variant of A*34 (e.g., A*34:01, A*34:02, etc.). At medium-resolution, only variants of A*34:02 (for example A*34:02:01G) are considered compatible, and any NMDP allele code containing the 34:02 allele.

We searched REDOME for full matching at three, four and five HLA loci, hereinafter referred to as 6/6, 8/8 and 10/10, respectively. Searches for 6/6 matching were for *HLA-A*, *-B* and *-DRB1*; 8/8 searches further include *HLA-C*; 10/10 searches further include *HLA-DQB1*. We also carried out analyses that allowed mismatches at a single allele (5/6, 7/8 and 9/10).

To evaluate the impact of the predicted accuracy of imputation on our results, we also performed searches on a subset of individuals with posterior probability of HLA imputation greater than 0.8 at all surveyed loci (2,554 individuals in total; REDS, n = 914; Bambuí, n = 433; Salvador, n = 219; Pelotas, n = 988). In the main text, all results refer to the dataset of 8,037 individuals, without filters for predicted accuracy of HLA imputation, and we refer to **Supplementary Materials** for results on the high confidence subset when appropriate.

## Inference of Genetic Ancestry

To infer the genetic ancestry for Brazilians, we used the three parental populations: 502 African and 503 European individuals from the 1000 Genomes Project phase-III (33), and 234 Native American individuals (34).

Genetic ancestry for each Brazilian individual was estimated with ADMIXTURE v.1.23 (35), which uses a maximum likelihood framework, based on multilocus SNP genotypes. We performed a supervised analysis with K=3 (corresponding to African, European and Native American ancestries), 2000 bootstrap replicates, windows of 50kb and step size of 10kb, and LD $R^2$ threshold of 0.1.

To infer the genetic ancestry of the MHC region for each individual, we used RFMix v.2 (36). First, genotypes for chromosome 6 were phased with the SHAPEIT v.2.12 software (37). Parental populations were subsampled to have equal sample sizes (234 individuals). We ran RFMix with default parameters, a time since admixture of 8 generations, and 2 EM iterations. Local ancestry was estimated for the whole of chromosome 6, but we subsequently focused on the ancestry of the subset of the MHC region encompassing the classical HLA loci, delimited by *HLA-A* and *-DQB1*, and spanning 2,724,220 bp.

## IBGE Categories, Genetic Ancestry and the Chance of REDOME Match

The Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística*, IBGE) defines a widely used classification, which identifies individuals as "Black" (*Preto*), "Mixed" (*Pardo*), "Indigenous" (*Indigena*), "White" (*Branco*) and "Yellow" (*Amarelo*), creating categories that confound skin color, social self-identification and genealogy (38). Despite critiques (39), this classification remains used in epidemiological studies, the Brazilian census, blood centers, and by REDOME. All individuals in the REDS-III and EPIGEN cohorts, as well as those in the REDOME database, are self-identified according to these categories. Here, we use only "Black", "Mixed" and "White", since "Indigenous" and "Yellow" contribute to only 1.8% of our sample (**Figures S2** and **S3**). Our use of these IBGE categories does not assume they are natural groupings with a biological basis. Rather, our goal is to examine how this widely used classification predicts the chances of finding a match in REDOME, and how results based on this classification compare to those obtained when groups defined by genetic ancestry are used.

We used a univariate logistic regression model, assuming "match in REDOME" (yes/no), as the outcome and IBGE categories or genetic ancestry as predictors, computing the odds ratio for each contrast. For genetic ancestry (estimated as percentages of African, European and Native American ancestries per individual), individuals were classified into quartiles of increasing African ancestry (Q1, Q2, Q3, Q4). For local ancestry analyses, we identified three genotypic classes for the MHC region in each individual: African/African, African/European or European/European. Because the ancestries in the MHC have a trimodal distribution (**Figure S5**), we could classify individuals into 3 groups according to number of chromosomes with African MHC: 0, 1 and 2 (rounding intermediate values, resulting from chromosomes with mixed ancestries, to the nearest integer).

# RESULTS

## IBGE Categories Explain Only a Small Amount of Genetic Ancestry

According to IBGE census (2019) (40) the Brazilian population has composition of 45.22% "White", 45.06% "Mixed", and 8.86% "Black", 0.47% "Yellow" and 0.38% "Indigenous", whereas REDOME's composition is 54.64% "White", 23.44% "Mixed", 7.17% "Black", 0.46% "Indigenous", 3.31% "Yellow" and 10.97% "Non Informed", indicating a deficit of the "Mixed" category in REDOME with respect to the Brazilian population as a whole.

The REDS-III and Salvador (EPIGEN) cohorts have high proportions of "Mixed" and "Black" and high African ancestry, whereas Bambuí and Pelotas are predominantly "White" and have high European ancestry (**Table 1**). The merged dataset has 44.1% "White", 32.6% "Mixed" and 19.8% "Black" individuals, and the average genetic composition was 62.4% European, 30.9% African and 6.1% Native American.

We initially examined the relationship between IBGE categories and genetic ancestry (**Figure 2**). Individuals categorized as "Black" have, on average, greater African genetic ancestry than those categorized as "White" (ANOVA p-value < $2x10^{-16}$; Tukey test p-value < 0.00001). Despite the statistical significance, much of the variation in ancestry among individuals is not captured by the IBGE categories, with about 36% of the variation in African and European genetic ancestry not being explained ($r^2$ = 0.64; **Figure S4**). The poor correlation between self-identified categories and genetic ancestry is even more apparent for the REDS and Salvador cohorts, both of which have a high proportion of individuals in the "Black" and "Mixed" categories (**Table 1**), and in which most of the variation in genetic ancestry is not captured by the IBGE categories ($r^2$ = 0.31).

**TABLE 1** | Brazilian Institute of Geography and Statistics (IBGE) categories and genetic ancestry across cohorts.*

| Dataset | IBGE categories (%) | | | | | | Genetic Ancestry (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Black | Indigenous | Mixed | White | Yellow | Non Informed | African | European | Native American |
| REDS-III | 26.7 | 0.5 | 58.6 | 10.7 | 0 | 3.5 | 49.7 | 43.6 | 6.7 |
| EPIGEN Salvador | 25.4 | 0.1 | 61.9 | 7.5 | 0 | 5.1 | 50.3 | 43.8 | 5.9 |
| EPIGENBambuí | 6.3 | 0 | 34.4 | 59.3 | 0 | 0 | 14.2 | 79.4 | 6.0 |
| EPIGEN Pelotas | 16.1 | 1.8 | 5.6 | 74.8 | 1.7 | 0 | 15.5 | 77.3 | 7.2 |
| Merged Dataset | 19.8 | 1 | 32.6 | 44.1 | 0.8 | 1.7 | 30.9 | 62.4 | 6.1 |

*Percentages refer to the proportion of individuals in each IBGE category, or the average per individual of each genetic ancestry.*

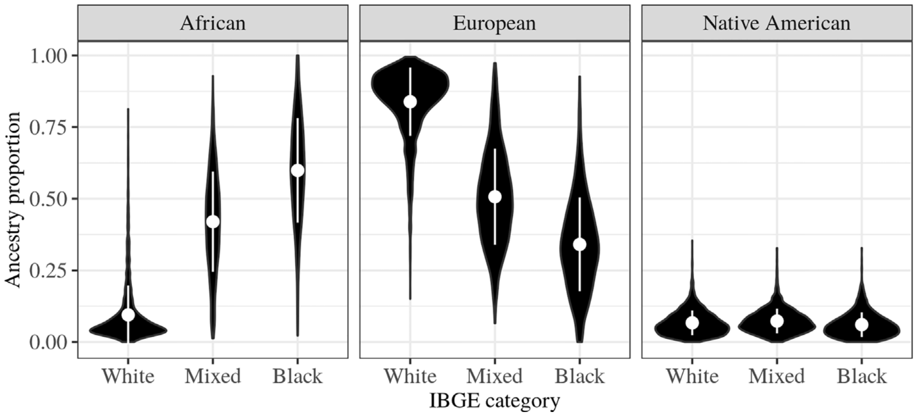


**FIGURE 2** | Relationship between Brazilian Institute of Geography and Statistics (IBGE) categories ("Black", "Mixed", and "White") and genetic ancestry (African, European and Native American) estimated using ADMIXTURE software in the merged dataset (REDS-III + EPIGEN). The white circles represent the averages and the vertical white lines correspond to one standard deviation, and the black shapes describe the distribution of ancestries within each IBGE category.

Our finding of a weak correlation between IBGE categories and genetic ancestry is consistent with previous studies of admixed Brazilians (18, 41), and suggests that the higher the admixture proportion, the lower the correlation between them. This suggests that there may be important differences between the influence of IBGE categories and of genetic ancestry on the chances of finding a match in REDOME, and it is this question we address in subsequent next sections.

## The Success of Finding Matches in REDOME

The proportion of individuals in the merged dataset for whom we find at least one match in REDOME at low resolution was 87.7% (6/6), 15.1% (8/8), and 13.0% (10/10). For medium resolution, matches were found for 51.3% (6/6), 6.1% (8/8), and 2.0% (10/10) of the individuals (**Table 2**).

The total number of potential donors per individual varies depending on the number of loci queried and resolution (**Table 2**). Among those individuals who find at least one match, the median number of potential donors at low resolution is 15 (6/6), 3 (8/8), and 3 (10/10); at medium resolution the values were 5 (6/6), 2 (8/8), and 3 (10/10).

Due to the small number of matches for medium resolution (and low statistical power to investigate the effect of ancestry), we subsequently only report results for low resolution matching. The results for all HLA combinations of low/medium resolution and numbers of loci are available in the **Supplementary Material** (**Figures S8–S16**).

## Self-Identification as “Black” and High African Ancestry Correlate With Lower Chances of Finding a Match

We first quantified the proportion of individuals with at least one match in REDOME (**Figure 3**, top row). In the 6/6 queries, 91.1% of the individuals classified as “White” find at least one donor compared to 84.7% and 82.9% of those categorized as “Mixed” and “Black”, respectively. In the 10/10 queries 16.9% of “White” individuals find a match, while only 7.3% of “Black” individuals do.

When we compare the chances of finding a match across groups defined by their proportion of African genetic ancestry, a similar pattern emerges, with the chance of finding a match decreasing as African genetic ancestry increases (**Figure 3**, middle row). For 6/6 queries, 92.1% of the individuals in the first quartile (with less than 0.61% African genetic ancestry) find at least one potential donor, as opposed to 81.7% of those in the fourth quartile (more than 51.31% African genetic ancestry). Again, in the 10/10 queries the difference is more extreme (18.6% vs. 7.12%).

While genetic ancestry refers to an average of the entire genome, it is possible to assign an ancestry specifically to the MHC region. We found that the greater the African ancestry in the MHC, the lower the chances of finding at least one potential donor. The percentage of individuals with 0 versus 2 chromosomes with African MHC who find a match is 93.0% vs. 76.5% at 6/6, and 19.2% vs. 4.4% at 10/10, respectively (**Figure 3**, bottom row).

Using a univariate logistic regression, all comparisons between groups with least African ancestry (i.e., “White” or Q1 of African ancestry) and most African ancestry (i.e., “Black” or Q4 of African Ancestry) are significant (p-value < $2\times10^{-16}$; see **Supplementary Material Tables S5** and **S6** for complete set of OR).

Among individuals with at least one match, the number of potential donors varies substantially (**Table 2**). We therefore assessed how the IBGE categories, genetic ancestry and MHC ancestry influence the number of potential donors found. Individuals categorized as “Black” and those with great African genetic ancestry, on average have a smaller number of potential donors, as compared to those categorized as “White” or having less African ancestry (**Figures S11–S13**). For the 6/6 queries, these differences are significant for all layers of ancestry (IBGE categories, African genetic ancestry, and MHC ancestry; p-value < 0.05, Mann-Whitney U test, **Table S7**).

Therefore, ancestry affects the chances of finding a match in two ways. First, in comparison with individuals categorized as “White” or having more European ancestry, a lower proportion of individuals classified as “Black” or carrying higher African ancestry find a match in REDOME (**Figure 3**). Second, among those who do find a donor, those classified as “Black” or having a greater African ancestry, on average have a smaller number of potential donors than “White” or genetically more European individuals (**Figures S11–S13**).

## Matching of Ancestries Between Donors and Recipients

Overall, our results support that ancestry influences the chances of finding a match in REDOME. Thus, we next asked whether the ancestry of an individual is the same as that of his/her potential donors. Since we do not have genetic ancestry data for REDOME, we analyze IBGE categories.

For individuals with at least one match, we computed the average proportion of potential donors from each IBGE

**TABLE 2** | Proportion of individuals with at least one match in Brazilian Bone Marrow Donor Registry (REDOME), median number of matches and the maximum number of matches at low and medium resolution.

| Loci | Low resolution | | | Medium resolution | | |
|---|---|---|---|---|---|---|
| | % at least one match | Median number of matches | Maximum number of matches | % at least one match | Median number of matches | Maximum number of matches |
| 6/6 | 87.7 | 15 | 3,157 | 51.3 | 5 | 2,229 |
| 8/8 | 15.1 | 3 | 81 | 6.1 | 2 | 56 |
| 10/10 | 13.0 | 3 | 79 | 2.0 | 3 | 49 |

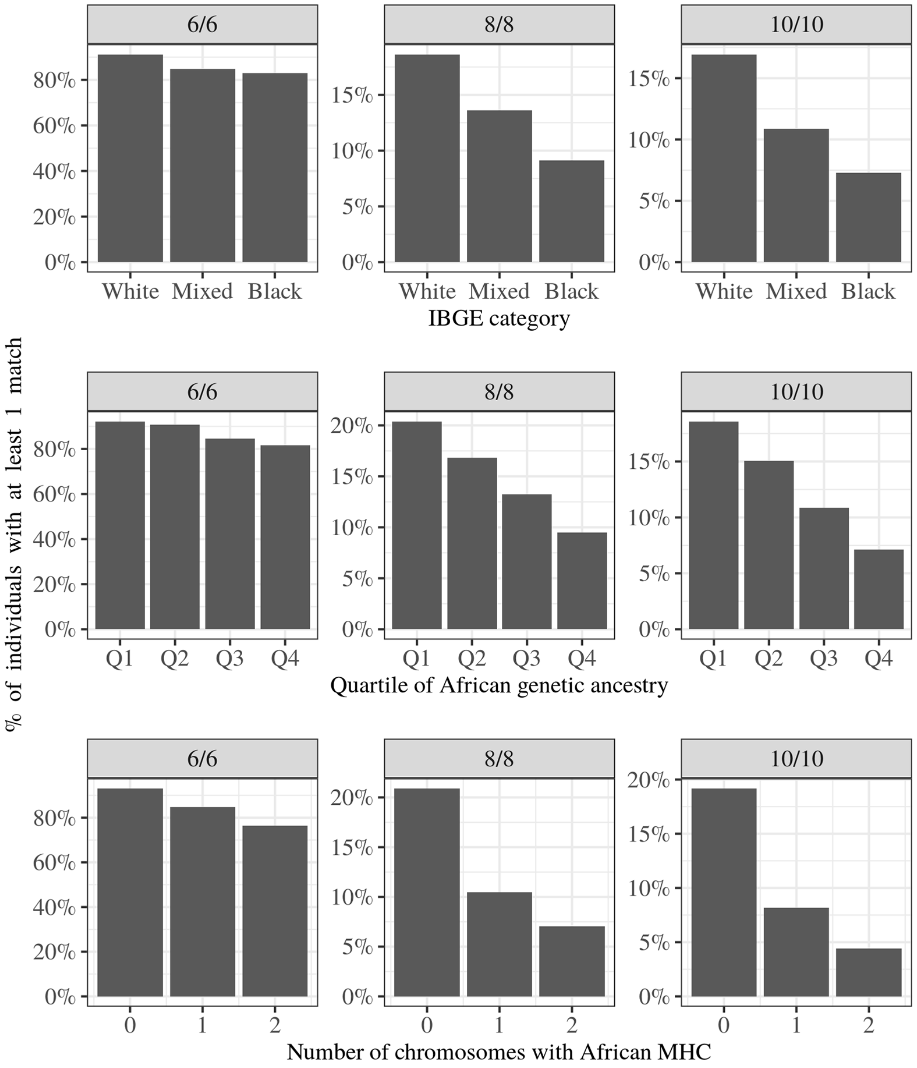


**FIGURE 3** | Percentage of individuals from the merged dataset with at least one match in Brazilian Bone Marrow Donor Registry (REDOME) for low resolution 6/6, 8/8, and 10/10. Top row: Brazilian Institute of Geography and Statistics (IBGE) categories; Middle row: Quartile of African ancestry; Bottom row: Number of chromosomes with African MHC per individual.

category. We find that individuals in the “Black” IBGE category, as compared to those in “White” and “Mixed”, match proportionally more to “Black” potential donors (5.7%, 10.1%, and 14.5% of “Black” donors for “White”, “Mixed” and “Black” individuals, respectively, at 6/6) (**Figure 4**, top row). When considering genetic ancestry, individuals with progressively more African ancestry match proportionately more to “Black” potential donors (4.8%, 6.4%, 10.4% and 15% of “Black” donors for individuals in Q1, Q2, Q3, Q4 at 6/6) (**Figure 4**, middle row). This trend is more pronounced when we consider ancestry in the MHC, where individuals who carry two chromosomes with African MHC match a pool of donors which is 20% “Black”, versus 4.5% for individuals with no chromosomes with African MHC (results for 6/6 matching) (**Figure 4**, bottom row).

## Finding Donors for Sickle Cell Disease Patients

Within our merged dataset, the individuals who are from the REDS-III cohort are diagnosed as having sickle cell disease (SCD). We therefore repeated our previous analyses on these individuals, so as to assess the influence of ancestry for a set of individuals who are eligible for HSTC (42).

Sickle Cell Disease (SCD) is a genetic disorder with an African origin, affecting 3,500 newborns annually, and estimated to affect between 25,000–30,000 people in Brazil (43). The only curative treatment available is HSCT. Although transplantation among family members is recommended, parents or siblings may be carriers of the sickle cell trait, which compromises them as donors. Thus, unrelated allo-HSCT transplantation or the haploidentical transplant, are useful options.

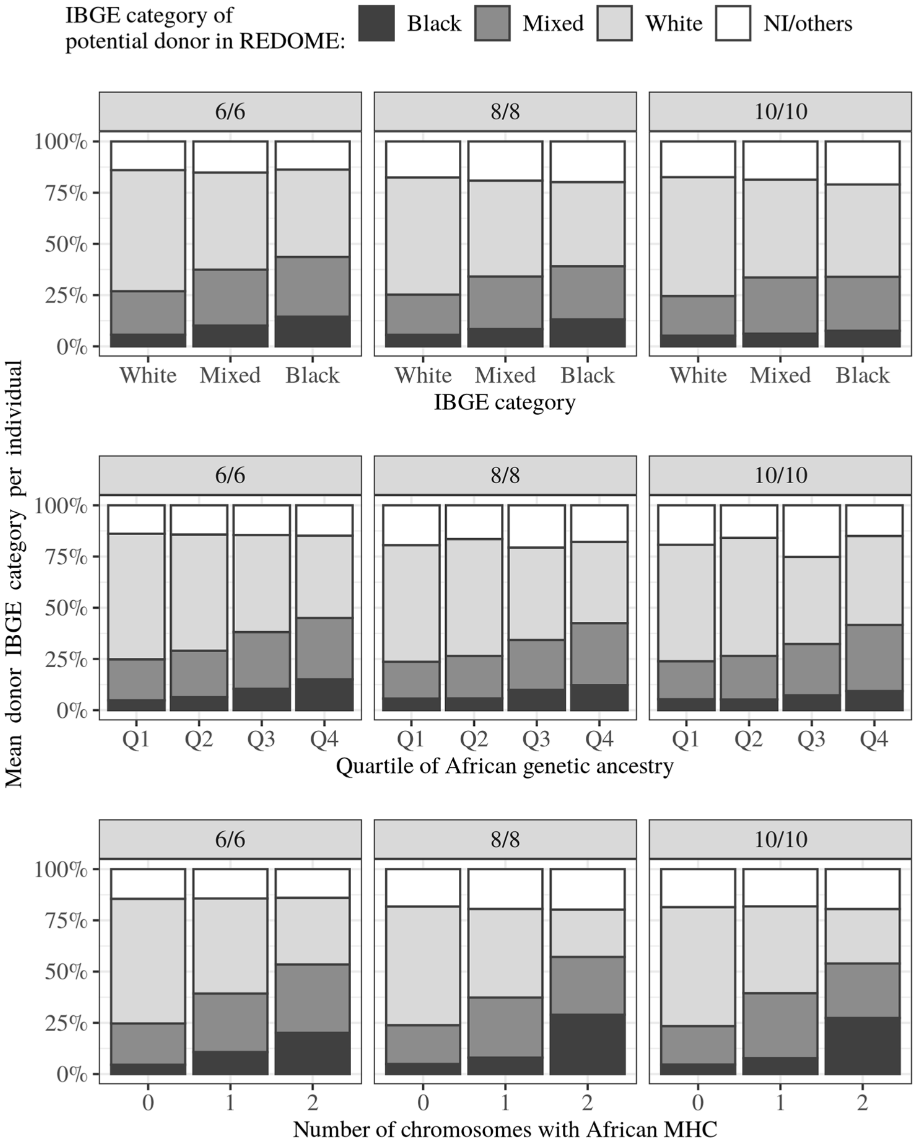


**FIGURE 4** | Mean frequency of Brazilian Institute of Geography and Statistics (IBGE) category of the potential donors for individuals in different IBGE categories, African ancestry quartiles, or with different numbers of chromosomes with African MHC. For each individual we calculate the average IBGE category of his/her potential donors, then for each category or genetic ancestry group we computed the group-level average (average of individual averages). Individuals who are "Black" or have more African genetic ancestry find proportionately more donors in the same category/ancestry.

Based on pre-established criteria by the Brazilian Ministry of Health, Flor-Park and col. (42) identified a subset of 417 patients within REDS-III who are eligible for transplantation. This subset consists of 30.3% "Black", 54.7% "Mixed", 10.3% "White" individuals, with 4.8% in other IBGE categories. The average genetic ancestry of these individuals is 52.0% African, 41.9% European and 6.1% Native American.

Among these SCD patients, 8.1% find a match in REDOME (10/10 low resolution), compared to 14.4% among individuals without SCD in the EPIGEN cohort (z-score=−3.563, p-value= 0.0004) (**Tables S9** and **S10**). Thus, consistent with the previous analyses, we show that individuals with increased African ancestry – in this case a set of patients with SCD and an indication for transplantation – have a decreased chance of finding a match in REDOME when compared to individuals from a group which has lower African ancestry (**Tables S8-S10**).

## DISCUSSION

We investigated how ancestry is related to an individual's chance of finding a match in REDOME, the third largest bone marrow registry in the world. For a sample of 8,037 Brazilians, we used three approaches to assess ancestry: (i) self-identified categories (e.g., "Black", "White", or "Mixed"), (ii) genetic ancestry (i.e., the percentage of an individual's genome which is African or European; **Figure S4A**), and (iii) genetic ancestry specific to the MHC region (i.e., how African or European are an individual's chromosomes within the MHC region; **Figure S4B**).

## A Hierarchy of Effects: Self-Identification, Genome-Wide Genetic Ancestry, and Genetic Ancestry in the MHC

As in other studies (17, 44, 45), we find that individuals categorized as "Black" or "Mixed" have a lower chance of finding a potential donor than "Whites". We also show that African genetic ancestry, both genome-wide and within the MHC, is associated with a marked reduction in the chances of finding a match. To directly compare the differences in chances of finding a match between classes such as "Black" and "White" to those between extremes of African ancestry, we performed an additional analysis, in which we matched the size of groups being compared (**Table 3**, explanation in legend). We document a hierarchy of effects of African ancestry on the chances of finding a match, with the chances progressively decreasing from IBGE categories, to genomic ancestry, and reaching a maximum decrease when comparing classes based on ancestry in the MHC region.

The association between African ancestry and decreased chances of finding a match has previously been explained as a consequence of the higher polymorphism of African populations (15). Humans originated in Africa, and spread over the world experiencing a series of bottlenecks (46, 47), which caused non-Africans to have lower genetic diversity. This reduced variation in non-Africans extends to HLA loci (9, 48, 49), and as a consequence the chance that two unrelated Africans will share a multi-locus HLA genotype is lower than that seen between two Europeans. In this study, we show that African ancestry in the MHC region is the strongest predictor of decreased chances of finding a match in REDOME. Our interpretation is that the IBGE categories and genetic ancestry genome-wide capture information about the ancestry within the MHC, thus accounting for their effects on the chances of finding a match. However, the ancestry within the MHC region is most directly associated with the chances of finding a match, since it captures information about the ancestry of the HLA allele an individual carries, which drives the chances of finding a match.

In addition to the role of African ancestry on the chances of finding a match, admixture in Brazilians may also be an important factor. The EPIGEN cohort (used in the present study), identifies that populations from different regions of the African continent contributed to the Brazilian gene pool (50), so that populations that never met in the African continent may experience gene flow in Brazil. If we transpose this issue to the MHC region, new HLA genotypic combinations may emerge among individuals of African ancestry, adding to the difficulty in finding unrelated voluntary donors. New genotypic combinations created by admixture may also be a more general issue, regardless of ancestry. For example, patients who do not find donors frequently carry alleles which are present in REDOME, but for which the genotype combination was not found (51). Thus, future studies should also model and quantify the role of admixture in shaping the chances of finding a match in REDOME.

**TABLE 3** | Percent decrease in the chances of finding at least one match for "Black" or most genetically African individuals, relative to "White" or least genetically Africans, respectively.*

| Query (low resolution) | Black vs. White | Most vs. Least African (genome) | Most vs. Least African (MHC) |
|---|---|---|---|
| 6/6 | 9% | 11% | 18% |
| 8/8 | 51% | 54% | 65% |
| 10/10 | 57% | 60% | 75% |

*All contrasts are referenced to that between "Black" (n=1,589) and "White" (n=3,544) sample sizes, involving a partition of ancestry of 3,544 individuals into two groups.*

## Clinical Implications

Our dataset contains a group of patients with sickle cell disease (SCD) and a recommendation for transplantation. At low resolution, 8.1% of SCD patients surveyed find a 10/10 match, compared to 14.4% for those without SCD (in the EPIGEN cohort; p-value=0.0004). This provides a concrete example of how patients with a higher than average African ancestry, when seeking transplantation, have reduced chances of finding a match.

There are still few studies with HSTC from unrelated donors for SCD patients [review in Oevermann and Sodani, 2020 (52)]. This scarcity is due to the difficulty of finding a match with unrelated donors in international public registries (52, 53). As a consequence, many of these unrelated allo-HSTC were performed with some level of HLA mismatch and were associated with a high prevalence of engraftment failure as well as graft-versus-host disease (53, 54). Recently, important advances have been reported, both with improved match between unrelated patients and donors (HLA high resolution full match or haploidentical) and in treatment with post-transplant drugs (55). However, these studies are not sufficient to guarantee the safety of this treatment (52), which is still evaluated as experimental. For this reason, HSCT between siblings is the gold standard for patients with SCD, and in Brazil it is the only modality recommended by the Ministry of Health (ordinance SAS/SCTIE n° 5/2018) with support from the Public Health System (Sistema Único de Saúde - SUS). Therefore, despite the curative potential of unrelated allo-HSCT for many hematological malignant diseases, it is not yet used routinely in clinical treatment of SCD. This is explained, to a large degree, by the high African genetic ancestry in patients with SCD, which decreases the chance of finding a compatible donor. As a consequence, the small number of transplants performed from unrelated donors have often allowed mismatches.

The drawbacks associated with the impossibility of locating HLA-compatible donors for HSCT are potentially severe. In the case of some hematological malignant diseases, such as acute lymphoblastic leukemia in adults, a lack of a HLA-compatible donor may mean withdrawing the therapeutic cornerstone, which is HSCT, and consequently compromising patient survival (56). This scenario is similar for other important non-malignant diseases, i.e. aplastic anemia, in which the HSCT plays a pivotal therapeutic role. However, it seems even more complicated in the case of SCD patients, for whom the low chance of finding a compatible unrelated donor makes it difficult to gather large samples to evaluate appropriate post-transplant parameters and treatment options. Therefore, at the moment, most SCD patients who are eligible for transplantation would not benefit from the positive effects of unrelated allo-HSCT, including decrease of vaso-occlusive episodes, donor-derived erythropoiesis and restoration/stabilization of function of damaged organs (57).

## HLA Diversity and Donor Recruitment Challenges

Previous studies have shown that individuals with the same ancestry tend to share HLA alleles more often than those from different populations (12, 58). Our study confirms this trend, with the number of "Black" donors increasing with the recipients' degree of African ancestry (**Figure 4**). On the other hand, many potential donors for "Black" individuals are in fact "White". This is expected in an admixed population for several reasons: many HLA alleles are shared around the world; REDOME is predominantly "White" and "Mixed"; self-identified categories do not accurately capture the genetic ancestry; potential "White" donors may contain African ancestry in the MHC region.

Faced with this complex scenario, it is natural to ask how large a registry should be, in order to adequately cover the diversity of a population. Expanding the size of the registry indefinitely is not an economically viable option, so strategies have been proposed to define an "optimal registry size". These efforts model the frequency of the most common HLA haplotypes, the distribution of individuals among regions within a country, and probability of matching (59–62). This task is particularly complex in admixed populations, since genetic diversity varies among groups.

As is the case for the NMDP (11, 14), REDOME has two features indicating that African ancestry is underrepresented, despite the registry's large size. First, there are proportionally fewer individuals classified as "Mixed" in REDOME as compared to the census for Brazil. Second, individuals with greater African ancestry have decreased chances of finding a match. Increasing REDOME's African ancestry component, as shown by our findings, is expected to increase the chance of individuals of African ancestry finding a match (**Figure 4**).

Once the underrepresented group has been identified, what is the best strategy for donor recruitment? It is economically unfeasible to determine the genetic ancestry of all donors. An alternative is to direct recruitment campaigns to regions/cities where the underrepresented group is more common, or to work with governmental and non-governmental organizations that represent these groups. The impact of a strategy directed to increase the presence of underrepresented groups in REDOME is likely to extend beyond Brazil, since between 2016-2018 REDOME exported 282 hematopoietic cells to Europe, the United States, and Asia (63).

## Caveats

In this study we resorted to imputation based on the SNP data to make HLA calls. The accuracy of imputation is reduced when the reference panel inadequately covers the diversity of the sample being imputed. Because the number of samples in available reference panels is still scarce, especially for admixed populations (64–67), alleles may have been incorrectly imputed. To evaluate the impact of imputation errors, we replicated our analyses for a subset of individuals with > 80% probability of correct imputation in all 5 HLA loci. In this dataset of HLA genotypes with a prediction of high accuracy, we find a greater proportion of individuals with at least one match (**Figure S17**, **Tables S11** and **S12**). This reflects the higher predicted accuracy of imputation for non-rare alleles and non-Africans, for which matching probabilities are also higher. However, the influence of ancestry on matching is highly concordant with the original conclusions, although some statistics are no longer significant (**Figure S17**, **Tables S11** and **S12**).

Our study was unable to explore the effects of Native American ancestry, which reaches 20% in Northern Brazil (39, 68). While a study of induced pluripotent stem cell (iPSC) banks in the USA showed similar match rates for Caucasians and Native Americans (48% and 46%, likely due to their low genetic diversities) (69), this remains to be explored in Brazil. However, the fact that most Native American alleles are present in REDOME (78% of endemic alleles were observed; https://genevol.ib.usp.br/wp-content/uploads/2020/05/AFND.html), and that their genetic diversity is very low, makes it likely that the finding of Pappas et al. (69) will also apply to Brazil, with match rates for Native Americans being similar to those of Europeans.

The nature of REDOME's data provides further challenges to our analyses. The data results from samples collected over a span of 30 years, resulting in substantial heterogeneity in the resolution of typing, and the number of loci typed per individual (**Table S4**). The samples collected earlier are at low resolution, whereas the most recent samples have the medium and high-resolution (including Next Generation Sequence data). As a consequence, our results for low and medium resolution, and across different numbers of loci, involve different sets of individuals and sample sizes. Given our focus on the role of ancestry and self-identification, we investigated whether there are differences in the distribution of IBGE categories across the data at different levels of resolution (or typed at a different number of loci). We found that the differences are very small (**Figure S18**), and therefore should not affect our findings.

Over most of the time of REDOME's operation, newly recruited voluntary donors were only typed for *HLA-A, -B* and *-DRB1*, with genotyping at *HLA-C* and *-DQB1*, as well as high resolution typing conditioned to an initial match at 6/6 (or 5/6). Throughout the paper, we analyze 6/6, 8/8 and 10/10 low resolution typing, despite the fact that successful transplantation requires full or partial matching at high resolution. Our results for medium resolution, 8/8 and 10/10 matching show that the effect of African ancestry becomes even stronger when additional loci are used. While the numbers presented cannot be used directly for clinical practice, the trends we identify are expected to hold in an analysis with high-resolution typing. We therefore believe that the true impact of ancestry on chances of finding a match may in fact be found to be even stronger than we report, when high resolution data for 5 loci become widely available in REDOME.

A final caveat refers to the fact that our study only queries the presence of a matching in REDOME, whereas the chance of finding a donor also involves the ability to contact the potential donors, their willingness to perform the transplant, confirmatory HLA typing at high resolution, and a medical evaluation of the donor's health. When these factors were taken into account, Gragert et al. (2014) (8) found that in the US donor availability for "Black" patients was 23%, compared to 51% for "White" patients. Since we do not have all this information for the Brazilian sample, we were unable to address these effects. However, the results of Gragert et al. (2014) (8) indicate that donor availability will be lower than the match rates we report, and that the effect of African ancestry on the chances of finding a match may be even greater than we document.

## Conclusion

We have shown that among patients seeking HSCT, those with a higher African ancestry face greater difficulty in locating potential donors. Given that more than half of the Brazilian population self-identifies as "Mixed" or "Black" [45.4% and 9.1%; (40)], and that these categories on average have increased African ancestry, this implies a reduced access to HSCT for a large proportion of Brazilians. Difficulty in obtaining a donor for HSTC is one among other forms of inequality faced by "Black" or "Mixed" Brazilians, who also carry a higher burden with respect to infant mortality (70), maternal mortality, risk of stroke (71), and more recently in the proportion of deaths due to COVID-19 (72). In these cases, self-identification is mainly a predictor of mortality as a consequence of the association to socio-economic status (71). In the case of access to donors in REDOME, we show that African ancestry is correlated with decreased chances of finding a match as an outcome of higher population-level diversity among African populations, associated with an underrepresentation of Brazilians of African ancestry in REDOME.

REDOME has a substantially higher frequency of "White" (54.6%) than "Mixed" or "Black" individuals (23.4% and 7.2%, respectively), implying a marked deficit of non-white categories in relation to that of Brazil as a whole. To illustrate a way to address this deficit, consider that individuals with two chromosomes with an African MHC have more matches to "Blacks" than to "Whites" (27.4% vs. 26.6%), even though "Blacks" represent only 7.2% of REDOME. This indicates that an increase in the proportion of individuals with African ancestry within REDOME will decrease the inequality in access to HSCT. This can be achieved, for example, by directing recruitment of new voluntary donors to regions where African ancestry is greater.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The data reported in this paper have been deposited in the European Nucleotide Archive European Nucleotide Archive (PRJEB26388 (ERP108374)), under EPIGEN Committee Controlled Access mode. And in the dbGAP (phs001972.v1.p1).

## ETHICS STATEMENTS

The studies involving human participants were reviewed and approved by Comissão Nacional de Ética em Pesquisa - Brazil (CONEP), the local ethics committees at each participating center, the Institutional Review Boards at the University of California, San Francisco, the REDS-III data coordinating center, RTI International, approved the study. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

KN, LP, and DM designed the study. KN, VA, MS, and AS analyzed the data. KN, VA, and DM wrote the manuscript. DO, CD, FK, ET-S, VR, AC-P, PL, MF-P, CM, SK, BC, BW, ES, and LP discussed, reviewed, and edited manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the United States National Institutes of Health - NIH (R01 GM075091; KN, VA, DM, BW); National Institutes of Health, National Heart, Lung, and Blood Institute (HHSN268201100007I; REDS-III); Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq and Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia e Insumos Estratégicos do Ministério da Saúde - Decit/SCTIE/MS (ET-S); Brazil Health Ministry – Transplant National System (proc. 25000.210075/2012-70; LP). São Paulo Funding Agency - FAPESP ( 2012/09950-9 KN and 2012/18010-0 DM). Minas Gerais State Research Agency (FAPEMIG RED-00314-16 ET-S).

## ACKNOWLEDGMENTS

This study was supported by the International Component of the NHLBI Recipient Epidemiology and Donor Evaluation Study (REDS-III). We gratefully acknowledge all of the research staff and investigators at each site for their commitment and contribution to this project: Fundação Pró-Sangue (São Paulo) – Alfredo Mendrone Jr., Cesar de Almeida Neto, Lígia Capuani; Instituto de Tratamento do Câncer Infantil (São Paulo) – Roberta Carlucci, Erivanda Bezerra; Hemominas–Belo Horizonte (Minas Gerais) – Franciane Mendes de Oliveira, Valquíria Reis, Nayara Duarte, Barbara Malta; Tassila Salomon, Carolina Miranda; Hemominas–Montes Claros (Minas Gerais) – José Wilson Sales, Maria Aparecida Souza, Rodrigo Ferreira, Rosi Afonso; Fundação Hemope–Recife (Pernambuco) – Aderson Araújo, Maria do Carmo Valgueiro, Regina Gomes, Airly Goes Maciel, Rebeca Talamatu Dantas; Hemorio – (Rio de Janeiro) – Flávia Herculano, Ana Claudia Pereira, Ana Carla Alvarenga, Adriana Grilo, Fabiana Canedo, Luiz Amorim; Instituto de Matemática e Estatística da Universidade de São Paulo – USP (São Paulo) – Pedro Losco Takecian, Rodrigo Muller de Carvalho, Mina Ozahata. US Investigators: RTI – Research Triangle Institute, International – Christopher McClure, Liliana Preiss, Don Brambilla; Vitalant Research Institute – Thelma Gonçalez; National Institutes of Health, National Heart, Lung, and Blood Institute – Simone A. Glynn; Instituto de Biociências da Universidade de São Paulo - Lilian Kimura; Registro Nacional de Doadores Voluntários de Medula Óssea - REDOME Brazil.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.584950/full#supplementary-material

## REFERENCES

1. Champlin RE, Gale RP. Role of bone marrow transplantation in the treatment of hematologic malignancies and solid tumors: critical review of syngeneic, autologous, and allogeneic transplants. *Cancer Treat Rep* (1984) 68:145–61.
2. Singh AK, McGuirk JP. Allogeneic Stem Cell Transplantation: A Historical and Scientific Overview. *Cancer Res* (2016) 76:6445–51. doi: 10.1158/0008-5472.CAN-16-1311
3. Tiercy J-M. How to select the best available related or unrelated donor of hematopoietic stem cells? *Haematologica* (2016) 101:680–7. doi: 10.3324/haematol.2015.141119
4. Duquesnoy RJ, Witvliet M, Doxiadis IIN, de Fijter H, Claas FHJ. HLAMatchmaker-based strategy to identify acceptable HLA class I mismatches for highly sensitized kidney transplant candidates. *Transpl Int* (2004) 17:22–30. doi: 10.1007/s00147-003-0641-z
5. Tiercy J-M. Unrelated hematopoietic stem cell donor matching probability and search algorithm. *Bone Marrow Res* (2012) 2012:695018. doi: 10.1155/2012/695018
6. Ruggeri A, Labopin M, Savani B, Paviglianiti A, Blaise D, Volt F, et al. Hematopoietic stem cell transplantation with unrelated cord blood or haploidentical donor grafts in adult patients with secondary acute myeloid leukemia, a comparative study from Eurocord and the ALWP EBMT. *Bone Marrow Transplant* (2019) 54:1987–94. doi: 10.1038/s41409-019-0582-5
7. Tiercy J-M, Claas F. Impact of HLA diversity on donor selection in organ and stem cell transplantation. *Hum Hered* (2013) 76:178–86. doi: 10.1159/000358798
8. Gragert L, Eapen M, Williams E, Freeman J, Spellman S, Baitty R, et al. HLA match likelihoods for hematopoietic stem-cell grafts in the U.S. registry. *N Engl J Med* (2014) 371:339–48. doi: 10.1056/NEJMsa1311707
9. Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol* (2008) 69:443–64. doi: 10.1016/j.humimm.2008.05.001
10. Fernandez Vina MA, Hollenbach JA, Lyke KE, Sztein MB, Maiers M, Klitz W, et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philos Trans R Soc Lond B Biol Sci* (2012) 367:820–9. doi: 10.1098/rstb.2011.0320
11. Bergstrom TC, Garratt R, Sheehan-Connor D. Stem Cell Donor Matching for Patients of Mixed Race. *B E J Econom Anal Policy* (2012) 12:746. doi: 10.1515/1935-1682.3275
12. Pidala J, Kim J, Schell M, Lee SJ, Hillgruber R, Nye V, et al. Race/ethnicity affects the probability of finding an HLA-A, -B, -C and -DRB1 allele-matched unrelated donor and likelihood of subsequent transplant utilization. *Bone Marrow Transplant* (2013) 48:346–50. doi: 10.1038/bmt.2012.150
13. Barker JN, Boughan K, Dahi PB, Devlin SM, Maloy MA, Naputo K, et al. Racial disparities in access to HLA-matched unrelated donor transplants: a prospective 1312-patient analysis. *Blood Adv* (2019) 3:939–44. doi: 10.1182/bloodadvances.2018028662
14. Bergstrom TC, Garratt RJ, Sheehan-Connor D. One Chance in a Million: Altruism and the Bone Marrow Registry. *Am Econ Rev* (2009) 99:1309–34. doi: 10.1257/aer.99.4.1309
15. Rosenberg NA, Kang JTL. Genetic Diversity and Societally Important Disparities. *Genetics* (2015) 201:1–12. doi: 10.1534/genetics.115.176750
16. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet* (2015) 96:37–53. doi: 10.1016/j.ajhg.2014.11.010
17. Hollenbach JA, Saperstein A, Albrecht M, Vierra-Green C, Parham P, Norman PJ, et al. Race, Ethnicity and Ancestry in Unrelated Transplant Matching for the National Marrow Donor Program: A Comparison of Multiple Forms of Self-Identification with Genetics. *PLoS One* (2015) 10: e0135960. doi: 10.1371/journal.pone.0135960
18. Lima-Costa MF, Rodrigues LC, Barreto ML, Gouveia M, Horta BL, Mambrini J, et al. Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci Rep* (2015) 5:9812. doi: 10.1038/srep09812
19. Sankar P, Cho MK. Genetics. Toward a new vocabulary of human genetic variation. *Science* (2002) 298:1337–8. doi: 10.1126/science.1074447
20. Gomes MB, Gabrielli AB, Santos DC, Pizarro MH, Barros BSV, Negrato CA, et al. Self-reported color-race and genomic ancestry in an admixed population: A contribution of a nationwide survey in patients with type 1 diabetes in Brazil. *Diabetes Res Clin Pract* (2018) 140:245–52. doi: 10.1016/j.diabres.2018.03.021
21. Santos DC, Porto LC, Oliveira RV, Secco D, Hanhoerderster L, Pizarro MH, et al. HLA class II genotyping of admixed Brazilian patients with type 1 diabetes according to self-reported color/race in a nationwide study. *Sci Rep* (2020) 10:6628. doi: 10.1038/s41598-020-63322-y
22. Carneiro-Proietti ABF, Kelly S, Miranda Teixeira C, Sabino EC, Alencar CS, Capuani L, et al. Clinical and genetic ancestry profile of a large multi-centre sickle cell disease cohort in Brazil. *Br J Haematol* (2018) 182:895–908. doi: 10.1111/bjh.15462
23. Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc Natl Acad Sci U S A* (2015) 112:8696–701. doi: 10.1073/pnas.1504447112
24. Barreto ML, Cunha SS, Alcântara-Neves N, Carvalho LP, Cruz AA, Stein RT, et al. Risk factors and immunological pathways for asthma and other allergic diseases in children: background and methodology of a longitudinal study in a large urban center in Northeastern Brazil (Salvador-SCAALA study). *BMC Pulm Med* (2006) 6:15. doi: 10.1186/1471-2466-6-15
25. Lima-Costa MF, Firmo JOA, Uchoa E. Cohort profile: the Bambui (Brazil) Cohort Study of Ageing. *Int J Epidemiol* (2011) 40:862–7. doi: 10.1093/ije/dyq143
26. Victora CG, Barros FC. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* (2006) 35:237–42. doi: 10.1093/ije/dyi290
27. Kuniholm MH, Xie X, Anastos K, Xue X, Reimers L, French AL, et al. Human leucocyte antigen class I and II imputation in a multiracial population. *Int J Immunogenet* (2016) 43:369–75. doi: 10.1111/iji.12292
28. Karnes JH, Shaffer CM, Bastarache L, Gaudieri S, Glazer AM, Steiner HE, et al. Comparison of HLA allelic imputation programs. *PLoS One* (2017) 12: e0172444. doi: 10.1371/journal.pone.0172444
29. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, et al. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. *Pharmacogenomics J* (2018) 18:367–76. doi: 10.1038/tpj.2017.7
30. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG–HLA genotype imputation with attribute bagging. *Pharmacogenomics J* (2014) 14:192–200. doi: 10.1038/tpj.2013.18

31. Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. *PLoS One* (2014) 9:e97282. doi: 10.1371/journal.pone.0097282

32. Abi-Rached L, Gouret P, Yeh J-H, Di Cristofaro J, Pontarotti P, Picard C, et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS One* (2018) 13:e0206512. doi: 10.1371/journal.pone.0206512

33. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* (2015) 526:68–74. doi: 10.1038/nature15393

34. Borda V, Alvim I, Aquino MM, Silva C, Soares-Souza GB, Leal TP, et al. The genetic structure and adaptation of Andean highlanders and Amazonian dwellers is influenced by the interplay between geography and culture. *bioRxiv* (2020) 174:1–17. doi: 10.1101/2020.01.30.916270

35. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* (2009) 19:1655–64. doi: 10.1101/gr.094052.109

36. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* (2013) 93:278–88. doi: 10.1016/j.ajhg.2013.06.020

37. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods* (2011) 9:179–81. doi: 10.1038/nmeth.1785

38. IBGE. *Brasil: 500 anos de povoamento. Centro de Documentação e Disseminação de Informações* (2007). Available at: https://biblioteca.ibge.gov.br/visualizacao/livros/liv6687.pdf (Accessed July 14, 2020).

39. Pena SDJ, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy F de SG, et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* (2011) 6:e17063. doi: 10.1371/journal.pone.0017063

40. População | IBGE. Available at: https://www.ibge.gov.br/estatisticas/sociais/populacao.html (Accessed July 14, 2020).

41. Lins TC, Vieira RG, Abreu BS, Gentil P, Moreno-Lima R, Oliveira RJ, et al. Genetic heterogeneity of self-reported ancestry groups in an admixed Brazilian population. *J Epidemiol* (2011) 21:240–5. doi: 10.2188/jea.je20100164

42. Flor-Park MV, Kelly S, Preiss L, Custer B, Carneiro-Proietti ABF, Araujo AS, et al. Identification and Characterization of Hematopoietic Stem Cell Transplant Candidates in a Sickle Cell Disease Cohort. *Biol Blood Marrow Transplant* (2019) 25:2103–9. doi: 10.1016/j.bbmt.2019.06.013

43. Cançado RD, Jesus JA. A doença falciforme no Brasil. *Rev Bras Hematol Hemoter* (2007) 29:204–6. doi: 10.1590/S1516-84842007000300002

44. Dehn J, Buck K, Maiers M, Confer D, Hartzman R, Kollman C, et al. 8/8 and 10/10 high-resolution match rate for the be the match unrelated donor registry. *Biol Blood Marrow Transplant* (2015) 21:137–41. doi: 10.1016/j.bbmt.2014.10.002

45. Bergstrom TC, Garratt R, Sheehan-Connor D. *Stem Cell Donor Matching for Patients of Mixed Race*. UC Santa Barbara: Department of Economics, UCSB (2009). Available at: https://escholarship.org/uc/item/22w466q9.

46. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U.S.A.* (2005) 102:15942–7. doi: 10.1073/pnas.0507611102

47. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. *Curr Biol* (2005) 15:R159–60. doi: 10.1016/j.cub.2005.02.038

48. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. *Genetics* (2006) 173:2121–42. doi: 10.1534/genetics.105.052837

49. Buhler S, Sanchez-Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PloS One* (2011) 6:e14643. doi: 10.1371/journal.pone.0014643

50. Gouveia MH, Borda V, Leal TP, Moreira RG, Bergen AW, Kehdy FSG, et al. Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. *Mol Biol Evol* (2020) 37:1647–56. doi: 10.1093/molbev/msaa033

51. Lopes RB. *Identificação de haplótipos HLA presentes em pacientes inscritos no REREME que não possuem doador compatível no REDOME*. [dissertation/master's thesis]. Rio de Janeiro (RJ): Universidade do Estado do Rio de Janeiro. (2018).

52. Oevermann L, Sodani P. Status quo of allogeneic stem cell transplantation for patients with sickle cell disease using matched unrelated donors. *Hematol Oncol Stem Cell Ther* (2020) 13:116–9. doi: 10.1016/j.hemonc.2019.12.004

53. Angelucci E, Matthes-Martin S, Baronciani D, Bernaudin F, Bonanomi S, Cappellini MD, et al. Hematopoietic stem cell transplantation in thalassemia major and sickle cell disease: indications and management recommendations from an international expert panel. *Haematologica* (2014) 99:811–20. doi: 10.3324/haematol.2013.099747

54. Shenoy S, Eapen M, Panepinto JA, Logan BR, Wu J, Abraham A, et al. A trial of unrelated donor marrow transplantation for children with severe sickle cell disease. *Blood* (2016) 128:2561–7. doi: 10.1182/blood-2016-05-715870

55. Gluckman E, de la Fuente J, Cappelli B, Scigliuolo GM, Volt F, Tozatto-Maio K, et al. The role of HLA matching in unrelated donor hematopoietic stem cell transplantation for sickle cell disease in Europe. *Bone Marrow Transplant* (2020) 55:1946–54. doi: 10.1038/s41409-020-0847-z

56. Pui C-H, Evans WE. Treatment of acute lymphoblastic leukemia. *N Engl J Med* (2006) 354:166–78. doi: 10.1056/NEJMra052603

57. Shenoy S. Hematopoietic stem-cell transplantation for sickle cell disease: current evidence and opinions. *Ther Adv Hematol* (2013) 4:335–44. doi: 10.1177/2040620713483063

58. Madbouly A, Wang T, Haagenson M, Paunic V, Vierra-Green C, Fleischhauer K, et al. Investigating the Association of Genetic Admixture and Donor/Recipient Genetic Disparity with Transplant Outcomes. *Biol Blood Marrow Transplant* (2017) 23:1029–37. doi: 10.1016/j.bbmt.2017.02.019

59. Sonnenberg FA, Eckman MH, Pauker SG. Bone marrow donor registries: the relation between registry size and probability of finding complete and partial matches. *Blood* (1989) 74:2569–78.

60. Schmidt AH, Sauter J, Pingel J, Ehninger G. Toward an optimal global stem cell donor recruitment strategy. *PloS One* (2014) 9:e86605. doi: 10.1371/journal.pone.0086605

61. Maiers M, Halagan M, Joshi S, Ballal HS, Jagannatthan L, Damodar S, et al. HLA match likelihoods for Indian patients seeking unrelated donor transplantation grafts: a population-based study. *Lancet Haematol* (2014) 1: e57–63. doi: 10.1016/S2352-3026(14)70021-3

62. Kwok J, Guo M, Yang W, Ip P, Chan GCF, Ho J, et al. Estimation of optimal donor number in Bone Marrow Donor Registry: Hong Kong's experience. *Hum Immunol* (2017) 78:610–3. doi: 10.1016/j.humimm.2017.08.007

63. Arrieta-Bolaños E, Oliveira DC, Barquera R. Differential admixture, human leukocyte antigen diversity, and hematopoietic cell transplantation in Latin America: challenges and opportunities. *Bone Marrow Transplant* (2020) 55:496–504. doi: 10.1038/s41409-019-0737-4

64. Levin AM, Adrianto I, Datta I, Iannuzzi MC, Trudeau S, McKeigue P, et al. Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. *BMC Genet* (2014) 15:72. doi: 10.1186/1471-2156-15-72

65. Nunes K, Piovezan B, Torres MA, Pontes GN, Kimura L, Carnavalli JEP, et al. Population variation of HLA genes in rural communities in Brazil, the Quilombos from the Vale do Ribeira, São Paulo - Brazil. *Hum Immunol* (2016) 77:447–8. doi: 10.1016/j.humimm.2016.04.007

66. Meyer D, Nunes K. HLA imputation, what is it good for? *Hum Immunol* (2017) 78:239–41. doi: 10.1016/j.humimm.2017.02.007

67. Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, et al. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. *Hum Mol Genet* (2019) 28:2078–92. doi: 10.1093/hmg/ddy443

68. de Souza AM, Resende SS, de Sousa TN, de Brito CFA. A systematic scoping review of the genetic ancestry of the Brazilian population. *Genet Mol Biol* (2019) 42:495–508. doi: 10.1590/1678-4685-GMB-2018-0076

69. Pappas DJ, Gourraud P-A, Le Gall C, Laurent J, Trounson A, DeWitt N, et al. Proceedings: human leukocyte antigen haplo-homozygous induced pluripotent stem cell haplobank modeled after the california population:

evaluating matching in a multiethnic and admixed population. *Stem Cells Transl Med* (2015) 4:413–8. doi: 10.5966/sctm.2015-0052

70. Matijasevich A, Victora CG, Barros AJD, Santos IS, Marco PL, Albernaz EP, et al. Widening ethnic disparities in infant mortality in southern Brazil: comparison of 3 birth cohorts. *Am J Public Health* (2008) 98:692–68. doi: 10.2105/AJPH.2006.093492
71. Chor D, Lima CR de A. [Epidemiologic aspects of racial inequalities in health in Brazil]. *Cad Saude Publica* (2005) 21:1586–94. doi: 10.1590/s0102-311x2005000500033
72. BOLETIM EPIDEMIOLÓGICO ESPECIAL. *Doença pelo Coronavirus COVID-19. Ministério da Saúde* (2020). Available at: http://saude.gov.br/images/pdf/2020/July/15/Boletim-epidemiologico-COVID-22.pdf (Accessed July 17, 2020).

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

ORIGINAL RESEARCH
published: 18 November 2020
doi: 10.3389/fimmu.2020.585731

# A Detailed View of KIR Haplotype Structures and Gene Families as Provided by a New Motif-Based Multiple Sequence Alignment

*David Roe*[1*], *Cynthia Vierra-Green*[2], *Chul-Woo Pyo*[3], *Daniel E. Geraghty*[3], *Stephen R. Spellman*[2], *Martin Maiers*[2] *and Rui Kuang*[1,4]

*[1] Bioinformatics and Computational Biology, University of Minnesota, Rochester, MN, United States, [2] Center for International Blood and Marrow Transplant Research, Minneapolis, MN, United States, [3] Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle WA, United States, [4] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, United States*

OPEN ACCESS

**Edited by:**
*Eleanor Riley,*
*University of Edinburgh,*
*United Kingdom*

**Reviewed by:**
*Hisashi Arase,*
*Osaka University, Japan*
*Ben Roediger,*
*Novartis Institutes for BioMedical*
*Research, Switzerland*
*Michael G. Brown,*
*University of Virginia, United States*

***Correspondence:**
*David Roe*
*roex0050@umn.edu*

**Specialty section:**
*This article was submitted to*
*NK and Innate Lymphoid*
*Cell Biology,*
*a section of the journal*
*Frontiers in Immunology*

**Received:** *21 July 2020*
**Accepted:** *20 October 2020*
**Published:** *18 November 2020*

**Citation:**
*Roe D, Vierra-Green C, Pyo C-W,*
*Geraghty DE, Spellman SR, Maiers M*
*and Kuang R (2020) A Detailed View of*
*KIR Haplotype Structures and Gene*
*Families as Provided by a New Motif-*
*Based Multiple Sequence Alignment.*
*Front. Immunol. 11:585731.*
*doi: 10.3389/fimmu.2020.585731*

Human chromosome 19q13.4 contains genes encoding killer-cell immunoglobulin-like receptors (KIR). Reported haplotype lengths range from 67 to 269 kb and contain 4 to 18 genes. The region has certain properties such as single nucleotide variation, structural variation, homology, and repetitive elements that make it hard to align accurately beyond single gene alleles. To the best of our knowledge, a multiple sequence alignment of KIR haplotypes has never been published or presented. Such an alignment would be useful to precisely define KIR haplotypes and loci, provide context for assigning alleles (especially fusion alleles) to genes, infer evolutionary history, impute alleles, interpret and predict co-expression, and generate markers. In order to extend the framework of KIR haplotype sequences in the human genome reference, 27 new sequences were generated including 24 haplotypes from 12 individuals of African American ancestry that were selected for genotypic diversity and novelty to the reference, to bring the total to 68 full length genomic KIR haplotype sequences. We leveraged these data and tools from our long-read KIR haplotype assembly algorithm to define and align KIR haplotypes at <5 kb resolution on average. We then used a standard alignment algorithm to refine that alignment down to single base resolution. This processing demonstrated that the high-level alignment recapitulates human-curated annotation of the human haplotypes as well as a chimpanzee haplotype. Further, assignments and alignments of gene alleles were consistent with their human curation in haplotype and allele databases. These results define KIR haplotypes as 14 loci containing 9 genes. The multiple sequence alignments have been applied in two software packages as probes to capture and annotate KIR haplotypes and as markers to genotype KIR from WGS.

**Keywords: killer-cell immunoglobulin-like receptors (KIR), alignment, DNA, haplotype, nomenclature, natural killer**

## INTRODUCTION

The proteins of the killer-cell immunoglobulin-like receptor (KIR) family are important to human health. Whether or not these transmembrane receptors on natural killer (NK) cells bind with peptide-bound human leukocyte antigen (HLA) class I molecules determines how they help educate, activate, and inhibit NK cell functionality, including cytotoxicity and cytokine release. The accumulated evidence seems to indicate that KIR evolved with HLA to balance its pathogen defense effects with its effects on reproduction *via* the embryo-uterus interface (1–3). KIR effects are generally tissue-specific and tissue-variable. Within a given NK cell, KIR expression is stochastic, depending on the epigenetic profiles, inter- and intra-genic content of the haplotypes, binding alleles or lack thereof, methylation status, alternative splicing, and randomness (4–6). Besides investigating KIR in the context of viruses and pregnancy, medicine is also studying its effects in cancer, hematopoietic stem cells transplants, various autoimmune diseases, and immunotherapy (7–9).

KIR gene names reflect their protein structures (10). First the prefix "KIR," followed by the number of extracellular domains ("2D" or "3D"), followed by a short or long intracellular domain ("S," L"), followed by an index. *KIR2DS1*, for example, has two extracellular domains ("2D") and a short intracellular domain ("S"); it is the first gene named with that structure ("1"), and *KIR2DS2* is the second. The Immuno Polymorphism Database for KIR (IPD-KIR) names KIR gene alleles, records their DNA-RNA-protein relationships, and annotates each gene's alleles in a multiple sequence alignment (11). It contains over 300 full-length DNA and almost 1,000 protein reference alleles in the latest release, 2.9.0.

There is no equivalent of IPD's allele annotation for haplotypes. KIR haplotypes have no official nomenclature, although the majority of contributions to the human genome reference use a convention set by Pyo *et al*. in 2010 (12) and 2013 (13). The haplotype names reflect the two-part structure of the region: a proximal centromeric ("c") region is paired with ("~") a distal telomeric ("t") region. Each region is a variant of a "A" haplotype or "B" haplotype family, followed by a two-digit index. The haplotype named "cB02~tA01," for example, is comprised of the second centromeric B region ("cB02") in cis with ("~") the first telomeric A region ("tA01").

Although there are many publications that analyze gene or intergene allele alignments, none report full haplotype multiple sequence alignments (MSAs), and therefore the haplotype nomenclature has not been formalized. Allele gene assignments are evaluated largely by amplification primers or sequence similarly to other alleles, as opposed to location in its haplotype. To help solve these issues, we present a simple bioinformatics approach to represent KIR haplotypes as a string of alignment-based motifs. We applied it by creating a full-haplotype DNA MSA for all 68 human haplotype sequences in the human genome project and a chimp haplotype for an outlier.

In our recent preprint describing a KIR long-read assembler (14), we show that 18 120 bp sequences can be used to capture full haplotype KIR DNA from PacBio circular consensus sequencing (CCS) reads. In this manuscript, we show when those probe sequences are aligned to assembled haplotypes, the pattern of the probes ("motifs") provides a structural annotation. These alignment locations of the motifs allow the haplotype sequences to be represented accurately and efficiently at the structural level. We leveraged that annotation to align the 68 published human KIR haplotypes and one chimpanzee that haplotype to within an average of 2,398 bases, and then we micro aligned each locus for precise full haplotype multiple sequence alignment. The results are concordant with the annotation in the human genome reference and reveal 13 structural haplotypes for the 68 human haplotype sequences. The MSA also shows the haplotypes have 14 loci containing 9 genes. *KIR2DS4* shares a locus with *KIR2DS3* and *KIR2DS5*, as opposed to *KIR2DS1* or by itself. Sharing a gene motif with locus *KIR2DS3*/*KIR2DS4*/*KIR2DS5* is the shared locus *KIR2DL2*/*KIR2DL3*. *KIR2DL1*, *KIR2DS1*, and *KIR2DS2* share a gene motif at different loci, as do *KIR2DL5A* and *KIR2DL5B*.

The MSA includes this study's contribution of 27 new haplotypes to the human genome project, including 24 haplotypes from 12 individuals of African American ancestry.

## MATERIALS AND METHODS

### Source Sequences and Annotation

The source sequences consisted of two newly assembled haplotypes from an Ashkenazim individual (from the assembler preprint), along with all 66 full-sequence alternative haplotypes in the human genome reference (12, 13, 15, 16) and a chimpanzee haplotype (GenBank accession AC155174.2) (17). The two Ashkenazim haplotypes were assembled with PacBio reads obtained from the Genome In a Bottle (GIAB) consortium (18) and have not yet been submitted to genetic databases. All other haplotypes were generated by physically separating chromosomes *via* fosmid cloning and then sequencing and assembling fragments into full haplotype sequences. The 68 include 27 new sequences first described in this manuscript, including 24 haplotypes from 12 individuals of African American ancestry that were selected by genotypic diversity and/or lack of representation in the human genome reference. An additional three haplotypes (one Asian, two European) of opportunity from two individuals were contributed by Scisco Genetics, who sequenced all non-GIAB haplotypes following previously reported protocols (12, 13). The GenBank entries also provided structural and allelic annotation of all the human haplotypes, except for the two Ashkenazim haplotypes. Broken down by population, the haplotype counts include 31 African or African American, 3 Asian, 2 Ashkenazim, 24 European, 2 Guarani South Amerindian, 2 Romani in Spain, and 4 unknowns. The structures are depicted schematically in **Figure 1**, excluding cA01~tB04, which contains a large insertion and was excluded for compact visualization.

### Workflow Overview

The result of the workflow was a multiple sequence alignment of 68 human and 1 chimp haplotypes. As shown in **Figure 2**, the first step was to align the 18 capture probes to the haplotype

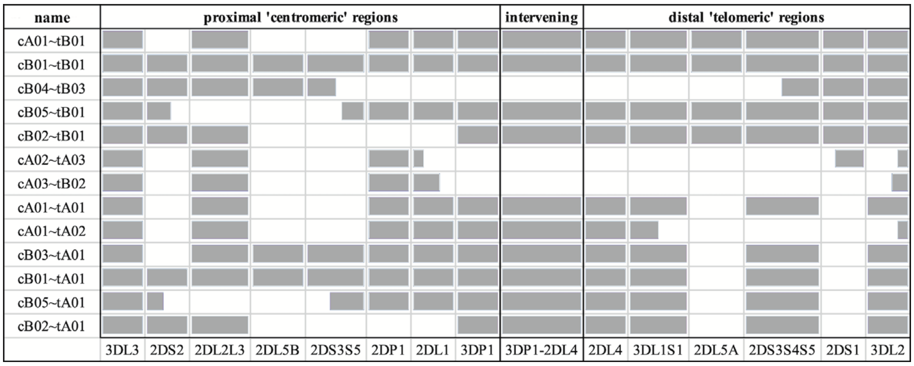


**FIGURE 1** | Schematic representation of the informal names and definitions of 13 human haplotype structures. The informal names of the haplotypes are in the first column. The definitions of the centromeric and telomeric regions are in the top row. The abbreviated names of the genes are in the bottom row; the intergenic region between *KIR3DP1* and *KIR2DL4* is represented as "3DP1-2DL4." Gray cells indicates the presence of alleles for that gene; white indicates absence of an allele. Some cells are partially colored to indicate two-gene fusion alleles. The names of the haplotypes, as well as the gene content and locations are taken from the human curated annotations in the human genome reference. Haplotype cA01~tB04 contains a large insertion and was excluded for visualization.

sequences (**Figure 2A**). Next, the sequence between each pair of probes (e.g., probe 4 to probe 3) was assigned a letter (e.g., C) (**Figure 2B**). In this way, each haplotype can be represented by a string ("motifs," e.g., "…CIJKL…") whose length is 105 characters or less. These haplotype motifs were aligned to produce a high-level MSA (**Figure 2C**). The motifs allowed loci to be located in the haplotype sequences. The sequences for all alleles of each locus was aligned separately, and then the MSAs were concatenated to create the full haplotype DNA MSA (**Figure 2D**). In this two-alignment approach, the motif alignment provides structural annotation and the locus-specific DNA alignment provides base-level accuracy. The following paragraphs detail each step.

## Probe Alignment and Inter-Probe Naming

The assembler preprint describes a method that uses 18 120 bp probes to capture KIR PacBio long-read sequences and assemble them into haplotypes with an average of 97% coverage and 99.7% concordance compared with reference sequences. When the probes are aligned to the haplotypes (**Figure 2A**), they align every 2,398 bases on average. Probe locations are discovered by alignment with bowtie2 with options "-a –end-to-end –rdg 3,3 –rfg 3,3." The alignment order of the 18 probes across a haplotype sequence allows that haplotype structure to be succinctly annotated as sequences of probe pairs (**Figure 2B**). For example, assume the alignment of probe 4 followed by probe 3 is called "C," probe 3 followed by probe 12 is called "I," and probe 12 followed by probe 10 is called "J." Then the region "probe 4 to probe 3 to probe 12 to probe 10" can be called "CIJ." Probe 1 repeating once in the alignment is "Z" and twice is "ZZ," etc. There are 42 such probe pairs in this collection of 69 haplotypes. In this way, haplotypes can be briefly annotated as strings of a 42-character alphabet. See **Figure 3** for an example of KP420440, whose full motif is MHCIJKLFGHCIJKLAIRLFZGHCIJ KLMHCIRLMHCIJKLSCTUVWXYKLSCNOJKLAIRLFZGHC IJKLMHCIJKLFPCNOJQ. The motif pairs are defined in **Supplementary Table 1**; the probe sequences are defined in the assembly manuscript and are also available *via* GitHub at https://github.com/droeatumn/kpi/tree/master/input.

## Inter-Probe Name Alignment

A full-haplotype multiple sequence alignment of the probe motifs was generated for the 68 human haplotypes plus the chimpanzee as an outgroup (**Figure 2C**). Alignment of the motifs was created with MAFFT (19), but mostly aligned manually with Aliview 1.2.6; manual alignment was required to resolve ambiguities, as the motif alphabet is not directly supported by DNA or protein aligners. Ambiguous alignments were resolved by following the human curation of the reference haplotypes.

## DNA Alignment

MAFFT was used to merge or add these haplotypes in the order cB02~tA01, cA01~tA01, cA02~tA01, cB02~tB01, cA0X~tB0X, cB0X~tA01, and cB0X~tB0X, where "X" a general number not already used. Then, the full-length DNA sequences were separated into sets as defined by their motif structures (**Figure 2D**). Each set was aligned separately with MAFFT. The alignment was then edited manually by adding or deleting gaps to conform to locations in the probe motif MSA. Using the loci defined in the motifs, and the motif locations defined in the DNA, the alignment was refined for all alleles in each locus with MUSCLE in Aliview. A high-level depiction of the alignment was created with Jalview 2.11.0's "Overview Window" functionality and NCBI's Multiple Sequence Alignment Viewer 1.13.1.

The MSA was validated at the structural level by showing that it recapitulates the human curation of the reference haplotypes in GenBank and the human genome reference. The alignment was validated at the allele level by showing each allele is assigned and annotated as expected with respect to IPD-KIR classifications.

The ability of existing software to align the 69 haplotypes was also evaluated with MAFFT stand-alone (–thread 19 –threadtb

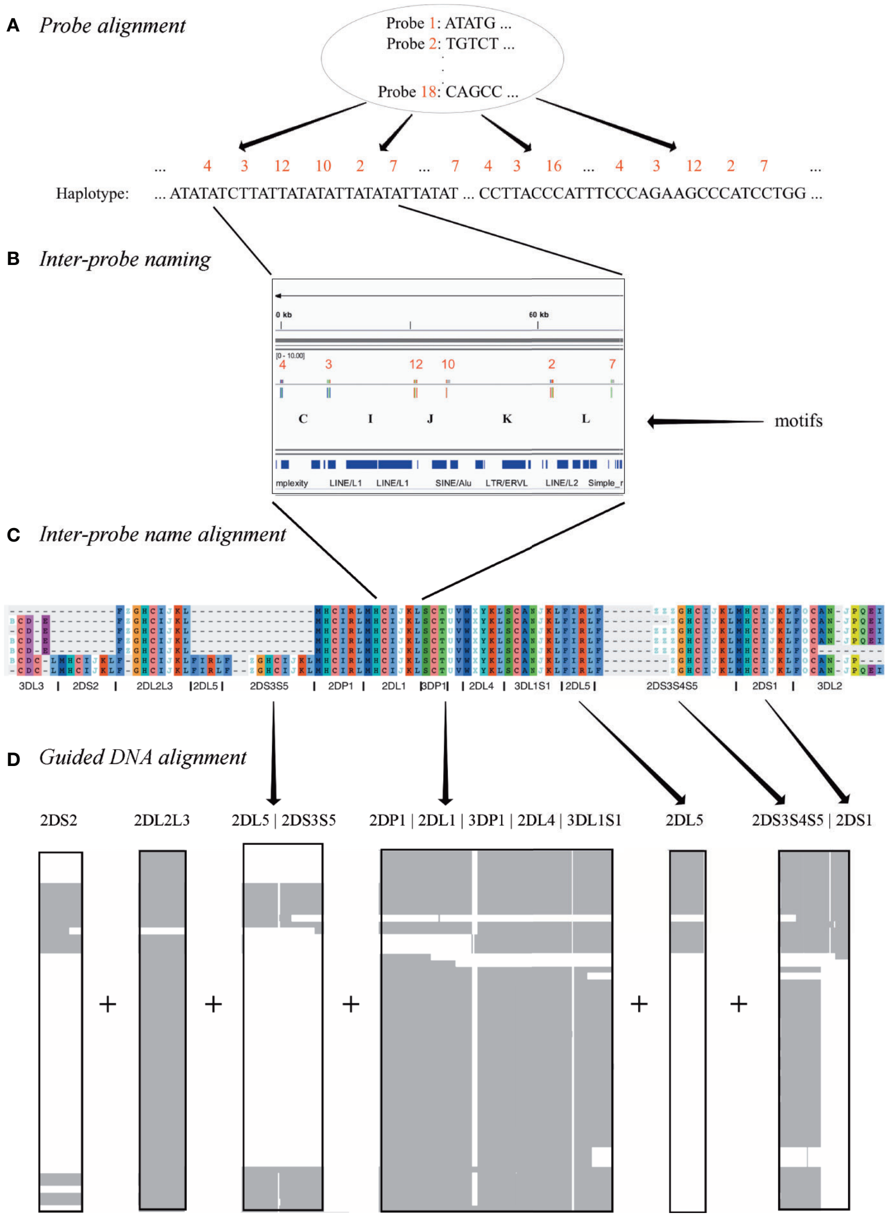

A Probe alignment
Probe 1: ATATG ...
Probe 2: TGTCT ...
Probe 18: CAGCC ...
... 4 3 12 10 2 7 ... 7 4 3 16 ... 4 3 12 2 7 ...
Haplotype: ... ATATATCTTATTATATATTATATATTATAT ... CCTTACCCATTTCCCAGAAGCCCATCCTGG ...
B Inter-probe naming
0 kb
60 kb
[0 - 10.00]
4 3 12 10 2 7
C I J K L
motifs
mplexity LINE/L1 LINE/L1 SINE/Alu LTR/ERVL LINE/L2 Simple_r
C Inter-probe name alignment
3DL3 | 2DS2 | 2DL2L3 |2DL5| 2DS3S5 | 2DP1 | 2DL1 |3DP1| |2DL4 | 3DL1S1 |2DL5| 2DS3S4S5 | 2DS1 | 3DL2
D Guided DNA alignment
2DS2
2DL2L3
2DL5 | 2DS3S5
2DP1 | 2DL1 | 3DP1 | 2DL4 | 3DL1S1
2DL5
2DS3S4S5 | 2DS1
+
+
+
+
+

**FIGURE 2** | Multiple sequence alignment (MSA) workflow. Step **(A)** depicts the alignment of 18 120 base probes to each haplotype sequence. Step **(B)** shows how each ordered pair of probes in the alignment is assigned a letter. For example, the letter "I" represents the sequence between the alignment of probe 3 and probe 12. Step **(C)** shows how each haplotype is depicted by a string of letters, how a haplotype motif MSA is generated from them, and how loci are defined in the MSA. Step **(D)** shows how the DNA in the motif-defined loci were separately aligned and then joined to create the final MSA.

10 –threadit 0 –reorder –adjustdirection –anysymbol –leavegappyregion –kimura 1 –auto), PASTA v1.8.5 (docker run –rm -it -v $PWD:/opt/droe smirarab/pasta run_pasta.py), Clustal Omega v1.2.4 (clustalo –threads=29), M-Coffee metaserver 13.41.0.28, MUSCLE v3.8.31 (default parameters), and WebPrank (Updated 8 October, 2017). Implementation was conducted on the web servers when possible or a server running Ubuntu 18.04.4 LTS with 32 AMD Opteron (TM) Processor 6220 and 200 GB RAM and maximum java heap space set to -Xmx100G.

# RESULTS

## Haplotype References

The African American haplotype sequences MN167504-MN167530 were deposited in GenBank; **Supplementary Table 2** contains details, including allele calls as IPD-KIR names for all 68 haplotypes in the human genome reference. Structural annotation for the new sequences was confirmed by the workflow by Pyo *et al.* for annotation of the previously-submitted reference haplotypes. The 27 haplotypes include 8 cA01~tA01, 4 cA01~tA02, 4 cB01~tA01,

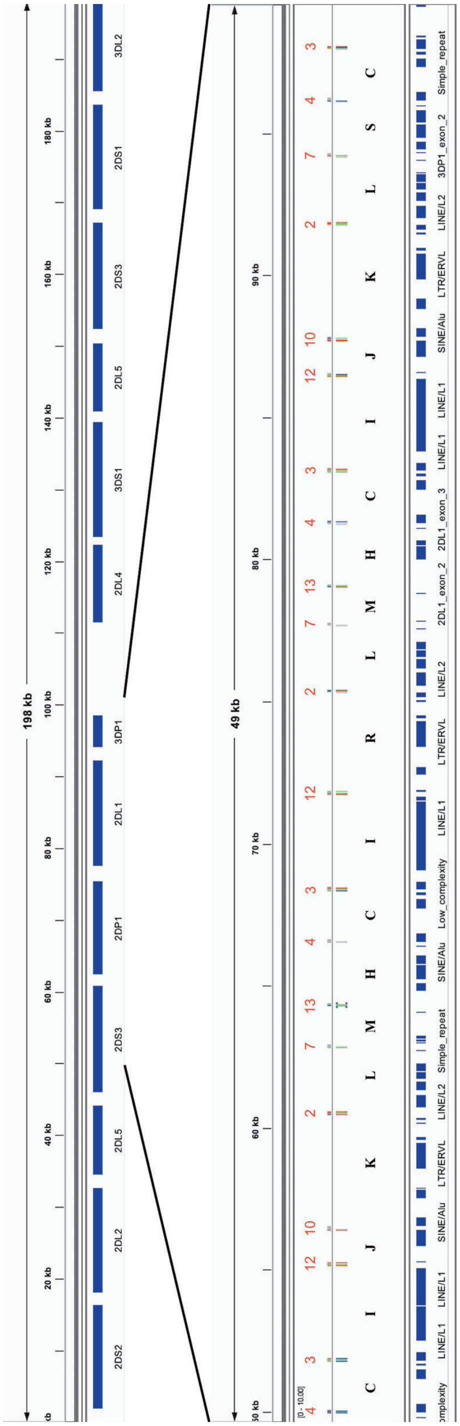


**FIGURE 3** | Example of the motif-generating alignment. The top frame shows the annotation for 198 kb haplotype KP420440 (cB01~tB01). The bottom frame zooms in to a 49 kb region depicting the alignment of the probes. In the middle of the bottom track, the locations of the probes are displayed by the vertical ticks with the red numbers above them and black letters below them. The red numbers label each probe, and the black letters label each probe pair (i.e., motif). The locations of exons and repeat elements (horizontal blue bars) are on the bottom track. Seven distinct probes align in this window. The motif pattern CIJKL occurs in two locations: 50–63kb and 81–94kb; the variation CIRL occurs in 66–79kb.

3 cB01~tB01, 3 cB03~tA01, 2 cA01~tB01, 2 cB02~tA01, and 1 cB01~tB01. Haplotype MN167506 (cB02~tB01), presumed to be of Asian ancestry, contains a *KIR2DL5B* allele (KIR2DL5B*00804) at the *KIR2DL5A* locus. Four haplotypes (MN167507, MN167509, MN167516, MN167530) with a *KIR3DL1*/*KIR3DL2* fusion (10) have been deposited in the human genome reference for the first time; they have been labeled as telomeric region "tA02." Some of the diversity of structures in the African American cohort are depicted visually in **Figure 4**. In total, the 68 human haplotype structures consist of 27 cA01~tA01, 7 cB01~tA01, 6 cB01~tB01, 6 cB02~tA01, 5 cA01~tB01, 4 cA01~tA02, 4 cB03~tA01, 3 cB02~tB01, and 1 each for cA01~tB04, cA02~tA03, cA03~tB02, cB04~tB03, cB05~tA01, cB05~tB01.

## Validation of the Multiple Sequence Alignment by Annotation in Human Genome Reference

**Figure 5** shows a multiple sequence alignment of the probe motifs for the human haplotypes, excluding insertion-containing cA01~tB04 for space considerations. It recapitulates the human-annotated structures in **Figure 1** with 105 positions. The common ~140 kb cA01~tA01 haplotypes are marked up in ~63 motif characters, and the ~220 kb cB01~tB01 haplotypes are marked up in ~94 motif characters. The average distance between probes is 2,398 bases; the maximum distance for non-*KIR3DL3* genes is ~5,700 bases and for *KIR3DL3* it is ~7,800 bases. The haplotype motifs subdivide into 14 genic and 1 intergenic (*KIR3DP1-KIR2DL4*) loci. From 5' (left) to 3' (right) at the bottom of **Figure 5**, the 15 abbreviated loci are: 3DL3, 2DS2, 2DL2L3, 2DL5, 2DS3S5, 2DP1, 2DL1, 3DP1, 3DP1-2DL4, 2DL4, 3DL1S1, 2DL5, 2DS3S4S5, 2DS1, and 3DL2. The 14 genic loci consist of 9 distinct motifs: 2DL1, 2DS1, 2DS2 share MHCIJ; 2DL5A and 2DL5B share FIRL; 2DS3, 2DS4, and 2DS5 share FZ +GHCIJKL. The genes can be summarized as a short motif, or in some cases a regular expression. For example, FZ+GHCIJKL means: a F followed by one-or-more Zs followed by GHCIJKL. **Supplementary Data Sheet 1** contains the motif MSA from **Figure 5**. **Supplementary Data Sheet 2** adds cA01~tB04 (KU645196) and the chimpanzee haplotype to the MSA. The chimp haplotype contains 55 motif characters, encoding genes that align to the human *KIR3DL3~KIR2DS2~KIR2DP1~KIR2DL1~KIR3DP1~KIR2DL4~KIR3DL1S1~KIR3DL2*.

Haplotype sizes by number of loci (not including the *KIR3DP1-KIR2DL4* intergenic region) range from four loci (cA04~tA03) to 18 (cA01~tB04). **Figure 6** displays a phylogenetic tree from the motifs of the 69 haplotypes, including cA01~tB04 and the chimp haplotype; the clusters recapitulate the their human-annotated names shown on the right part of the figure.

The DNA alignment of the 67 human haplotypes (minus cA01~tB04) has 263,556 positions; it is included in **Supplementary Data Sheet 3**. **Figure 7B** shows an overview of the DNA haplotype alignment as generated by Jalview's "Overview Window" function; the names of the haplotypes were added after exporting the overview image from Jalview. The pattern of white (gaps) and gray (aligned sequences) mirror the patterns in the schematic depictions of the human curated

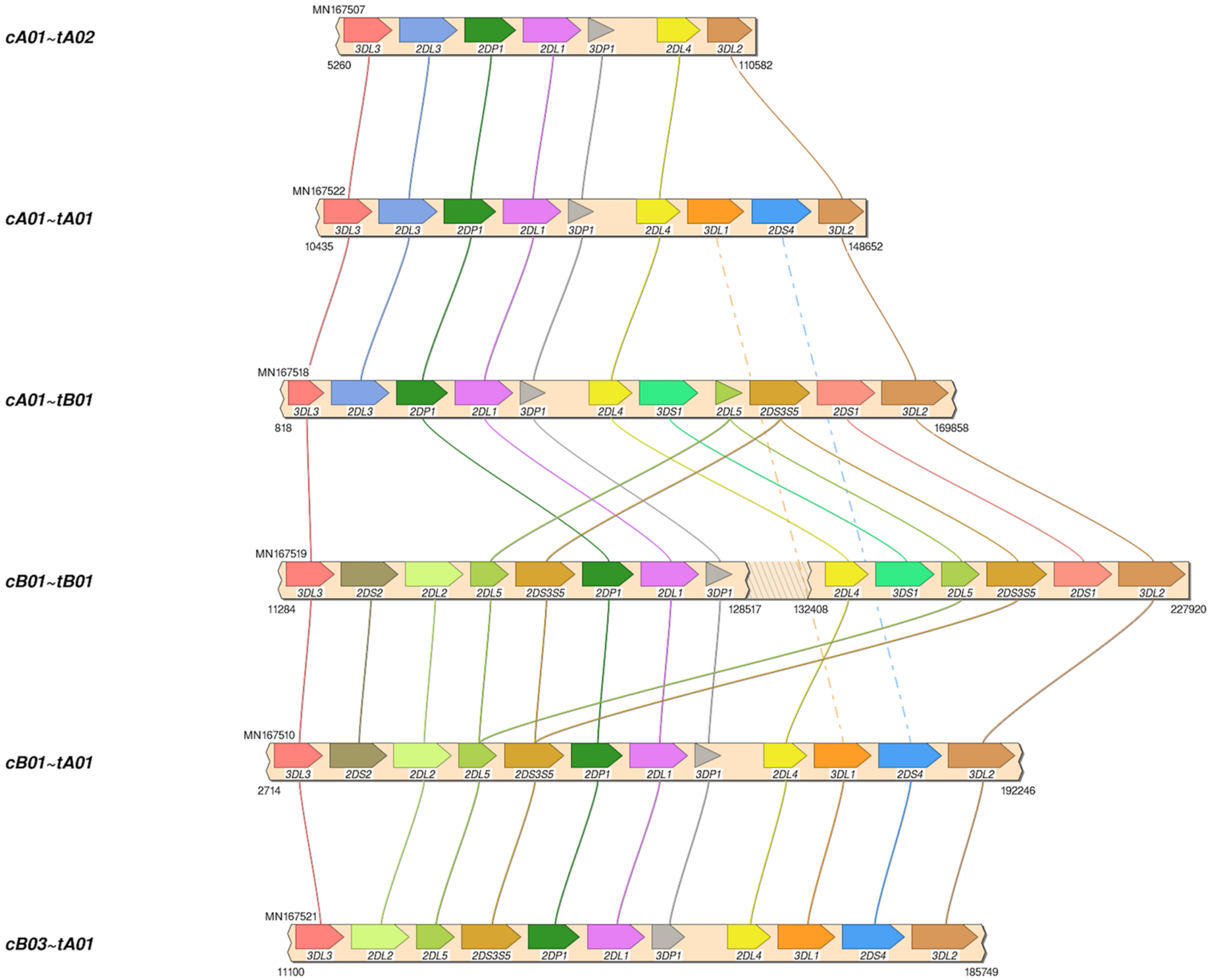

cA01~tA02
MN167507
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DL2
5260
110582
cA01~tA01
MN167522
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DL1 2DS4 3DL2
10435
148652
cA01~tB01
MN167518
3DL3 2DL3 2DP1 2DL1 3DP1 2DL4 3DS1 2DL5 2DS3S5 2DS1 3DL2
818
169858
cB01~tB01
MN167519
3DL3 2DS2 2DL2 2DL5 2DS3S5 2DP1 2DL1 3DP1 2DL4 3DS1 2DL5 2DS3S5 2DS1 3DL2
11284
128517
132408
227920
cB01~tA01
MN167510
3DL3 2DS2 2DL2 2DL5 2DS3S5 2DP1 2DL1 3DP1 2DL4 3DL1 2DS4 3DL2
2714
192246
cB03~tA01
MN167521
3DL3 2DL2 2DL5 2DS3S5 2DP1 2DL1 3DP1 2DL4 3DL1 2DS4 3DL2
11100
185749

**FIGURE 4** | Haplotype structures in African American cohort. Shown are six of the eight structures in the African American cohort, which was added to the human genome project as part of this study. Not shown are cB02~tA01, and 1 cB01~tB01.

haplotypes in **Figure 1** (recapitulated in **Figure 7A**). This evidence shows the DNA alignment matches the human curation at a structural level.

**Figure 8** shows a view of the alignment of the 67 haplotypes using NCBI's MSA Viewer. Positions that are in an agreement with consensus are colored in gray, positions that are not in an agreement with consensus are colored in red. In this tool, consensus includes gaps (lack of an allele at that position). For KIR, this causes the regions specific to B haplotypes to show as pure red, and the other regions to show as mixture of red and gray.

## Validation of the Multiple Sequence Alignment-Defined Allele Assignments by Immuno Polymorphism Database for KIR Annotation

The haplotypes are annotated by the names and orders of their alleles in **Supplementary Table 2**. Of the 647 alleles in the 68 haplotypes, 556 (86%) can be assigned names *via* IPD-KIR 2.9.0, at least at protein resolution. Unnamed alleles occur either because they have not yet been included in the IPD-KIR database, or they are partial alleles in the case of *KIR3DL3* and *KIR3DL2*. 97% of alleles can be named excluding *KIR3DL3*, *KIR2DP1*, and *KIR3DL2*. 39 *KIR2DP1* alleles are unnamed. To check the gene assignment for the 91 alleles that could not be named, the sequences were aligned to a set of 17 full gene alleles, each the first named allele for its assigned gene in IPD-KIR. Each allele being evaluated was considered to be correctly annotated if the allele sequence aligned closet to the IPD-KIR reference to which it was assigned by the motifs. Those results show that every unnamed allele aligned to the reference allele predicted by its motif assignment. The only exception was the GU182360 *KIR3DL2*, which aligns closest to *KIR3DP1*; however, this haplotype sequence is incomplete on the 3' end, and the *KIR3DL2* allele only contains the 2,221 5'-end sequences. Alleles that are a fusion of *KIR3DL1* and *KIR3DL2* (e.g., *KIR3DL1* alleles 059-061) are classified as *KIR3DL2* alleles using motifs but are classified as *KIR3DL1* alleles in IPD-KIR.

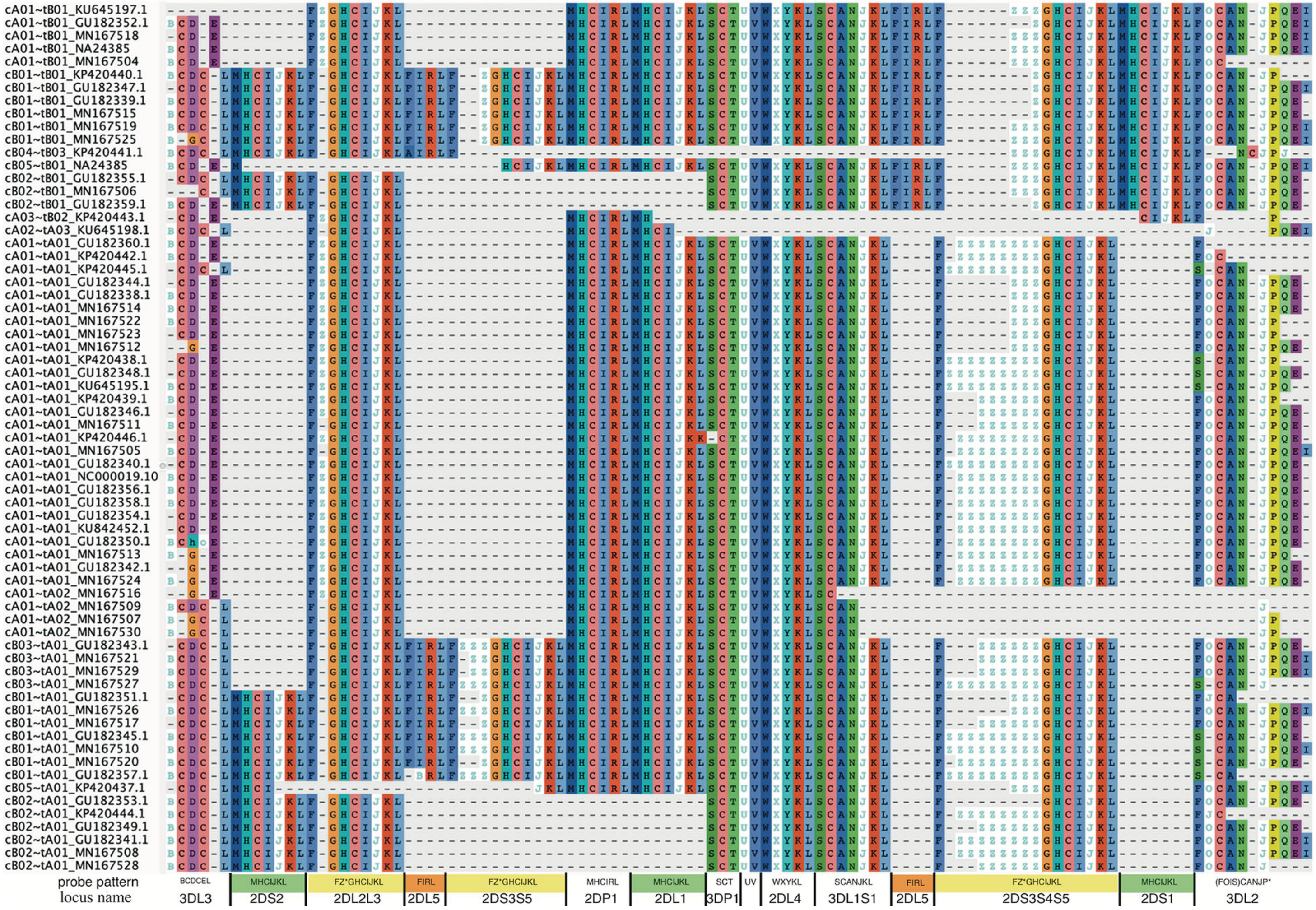


**FIGURE 5** | Probe motif multiple sequence alignment of the human haplotypes. The informal haplotype name and GenBank accession number are in the first column. The consensus motif pattern with each abbreviated locus name is described in the bottom row. Some loci share motifs: *KIR2DL1*, *KIR2DS1*, *KIR2DS2* (green); *KIR2DL5A* and *KIR2DL5B* (orange); *KIR2DS3*, *KIR2DS4*, and *KIR2DS5* (yellow). Some gene names are combined when they align at the same locus: *KIR2DL2* and *KIR2DL3* (2DL2L3), *KIR2DS3* and *KIR2DS5* centromeric side (2DS3S5), *KIR2DS3*, and *KIR2DS4*, and *KIR2DS5* telomeric side (2DS3S4S5), and *KIR3DL1* and *KIR3DS1* (3DL1S1). Haplotype cA01~tB04, which contains a 106 kb insertion, was omitted for visualization.

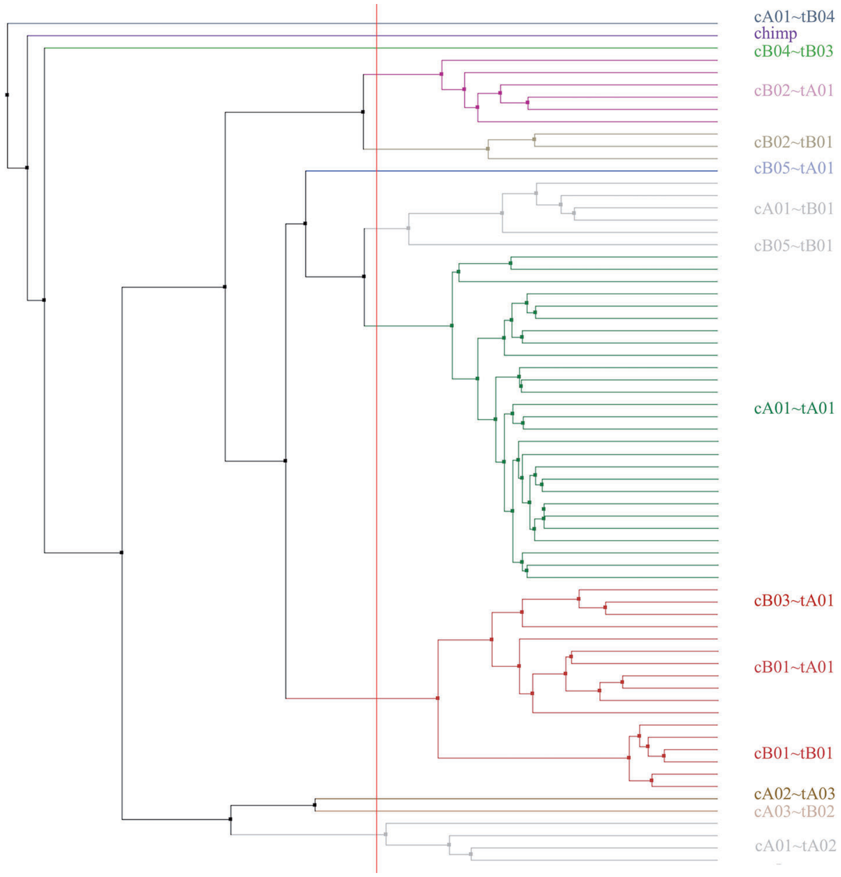


**FIGURE 6** | Phylogenetic tree made from the multiple sequence alignment (MSA) from the motifs of all the human haplotypes plus the chimpanzee (n=69). The leaf descriptions on the right include the informal name of the structural haplotype as documented in GenBank. The leaves are colored by their alignment order and the grouping defined by the red vertical line. The tree may not reflect the evolutionary history.

## Comparison with Existing Multiple Sequence Alignment Software

Of the existing alignment software that was evaluated on the 67 human haplotypes, only stand-alone MAFFT was able to generate an alignment, using the FFT-NS-2 strategy. The input size was too large for Clustal Omega local server, M-Coffee web server, MUSCLE local server, and WebPrank web server, and the other algorithms ran out of memory. The MAFFT alignment, output, and overview are included in **Supplementary Data Sheet 4** and **Supplementary Image 2**. Except for large portions of *KIR3DL3*, distal *KIR2DL3*/*KIR2DS4*/2DS3S5, and *KIR3DL2*, the MAFFT alignment does not recapitulate human-curated KIR haplotypes or gene alleles. It is 75% larger than our motif-guided alignment (960,221 vs. 263,556 positions), and its overview shows the alignment columns are mostly gaps compared with the more expected block shaped column in **Figure 7**.

# DISCUSSION

Although our evaluation of existing MSA software methods was not exhaustive, we believe it is unlikely that any current general-purpose alignment software can align all human KIR haplotype sequences consistent with the human curation. Alignment of the 68 sequences, each 67–269 kb, each homologous with itself and every other haplotype over lengths of 15 kb is very challenging without prior knowledge. Scoring matrices and algorithms like dynamic programming do not generally support sequences of this size under conditions where relatively young gene regions have duplicated both within and between haplotypes. To the best of our knowledge, a MSA of KIR haplotypes has never been published, despite the fact that dozens of haplotype sequences have been public for years.

**Figures 5** and **6** demonstrate that the capture probe alignment motifs alone can annotate KIR haplotype sequences in a way that recreates the human curated annotation in GenBank. **Figures 7** and **8** similarly show the accuracy of the DNA alignment; they demonstrate that the DNA annotated by the motifs recreates the assignment of the alleles in IPD-KIR. Also, for two practical examples, the capture probes have demonstrated the ability to assemble haplotypes with an average of 97% coverage (preprint), and the DNA MSA was used to discover gene markers that can genotype whole genome

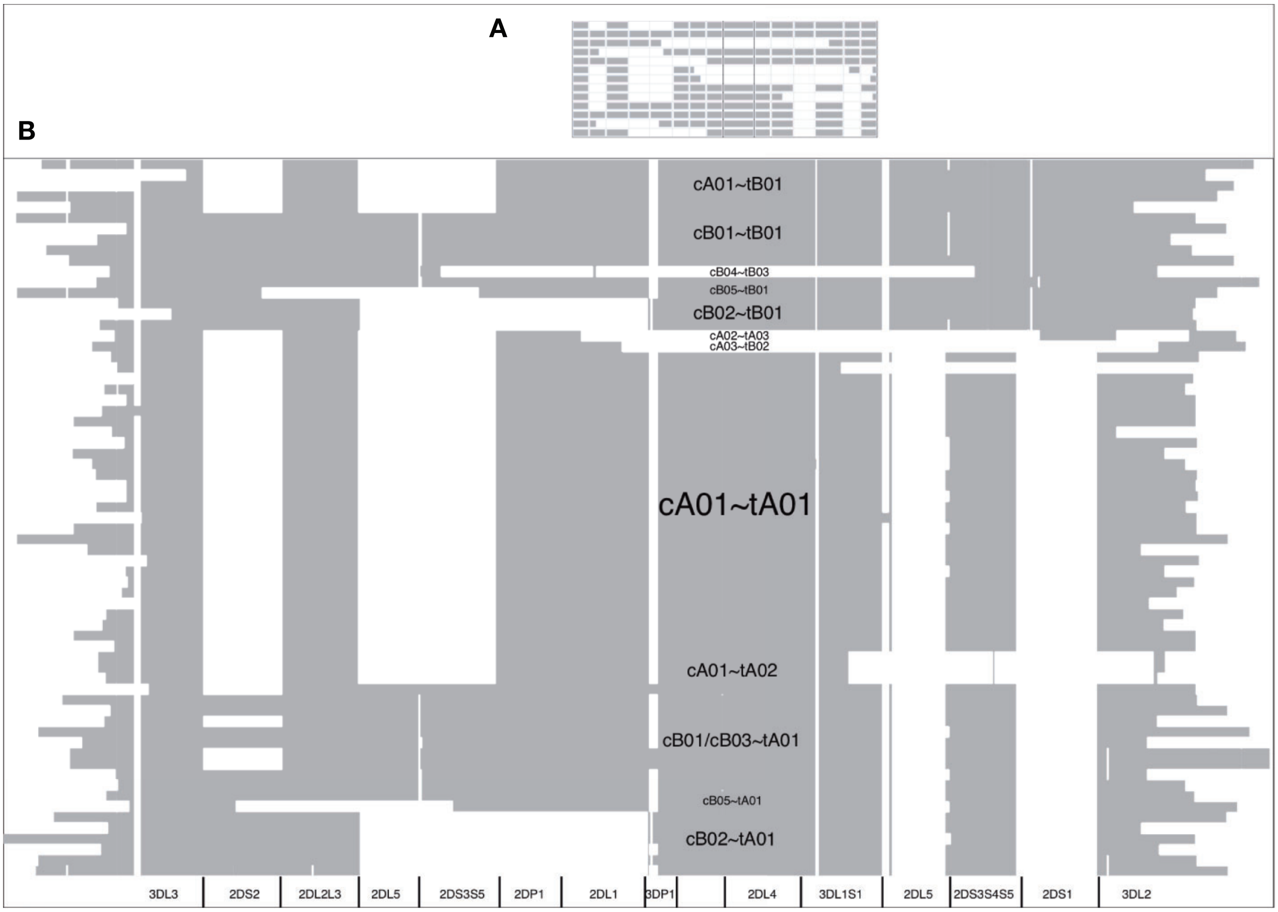


**FIGURE 7** | DNA multiple sequence alignment (MSA) overview. The cartoon of the GenBank annotations from **Figure 1** is in **(A)** and is included to compare with the Jalview overview of the DNA full-haplotype MSA from the 67 haplotypes in **(B).** The gray denotes alignment of the DNA bases. White denotes gaps between the sequences. cA01~tB04 is excluded for better visualization.

sequences (WGS) with almost perfect accuracy (11) (preprint). The combination of multi-resolution validation and software applications demonstrates that both the motif and DNA alignments are likely to be accurate.

**Figure 5** (alignment in **Supplementary Data Sheet 1** suggest 14 human KIR loci and 1 intergenic locus. The genic loci are abbreviated and labeled as 3DL3, 2DS2, 2DL2L3, 2DL5B, 2DS3S5, 2DP1, 2DL1, 3DP1, 2DL4, 3DL1S1, 2DS3S4S5, 2DS1, 3DL2, and the intergenic locus is between 3DP1 and 2DL4 (3DP1-2DL4). There are nine motif-defined genes in the 14 genic loci (**Figure 5**). Loci abbreviated 2DL1, 2DS1, and 2DS2 share motif MHCIJKL. Loci 2DL2L3, 2DS3S5, and 2DS3S4S5 share variations of the FZ*GHCIJKL motif. Loci 2DL5A and 2DL5B share variations of FIRL. The motif MSA suggests *KIR2DS4* on the A haplotype shares a locus with *KIR2DS3* and *KIR2DS5* (not *KIR2DS1*) on the B haplotypes. The MSA similarly suggests *KIR2DL2* and *KIR2DL3* share a locus as do *KIR3DL1* and *KIR3DS1*. The Z character is unique to centromeric 2DS3S5 and telomeric 2DS3S4S5 loci and marks the alleles to a certain extent. In both, the *KIR2DS3* alleles have 1 Z in their motifs, *KIR2DS5* has 2 or 3, and *KIR2DS4* alleles are more variable.

Although the motifs correctly annotate the haplotype structures, they do not always do so unambiguously, nor do they always agree with the existing nomenclature as utilized in IPD-KIR. In the current gene nomenclature, there are 16 KIR genes. Some researchers have previously considered *KIR2DL2* and *KIR2DL3* to occupy the same locus as well as *KIR2DS3* and *KIR2DS5* (20, 21). Conversely, *KIR2DL5* was once considered one gene but was split into *KIR2DL5A* for alleles in the centromeric locus and *KIR2DL5B* for alleles in the telomeric locus, creating new “A” and “B” designations that are independent from the “A” and “B” haplotype classifications (22). The motif MSA suggests the KIR region has 14 loci in 9 gene motifs. Under the existing gene nomenclature, *KIR3DL1* and *KIR3DL2* differ only by an index, since they are both three domain (“3D”) long tail (“L”) genes; since the extracellular domains of *KIR3DL1/KIR3DL2* fusion alleles are from *KIR3DL1*, those fusion alleles are labelled in IPD-KIR as *KIR3DL1*. They are considered *KIR3DL2* alleles under the motif

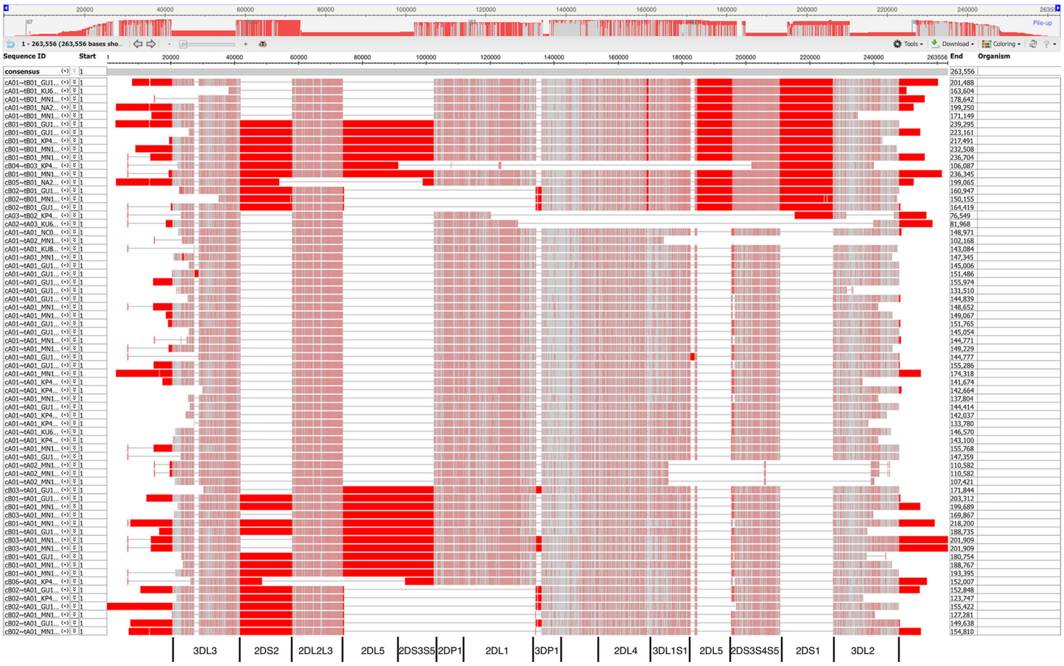


**FIGURE 8** | Multiple sequence alignment (MSA) Viewer depiction of the DNA alignment of 67 human haplotypes. Positions that are in an agreement with consensus are colored in gray, positions that are not in an agreement with consensus are colored in red. The location and borders of the genes have been added at the bottom.

structure because the proximal portions of *KIR3DL1* and *KIR3DL2* share variations of SCANJ, but the distal portions are different; since the fusions are comprised of proximal *KIR3DL1* and distal *KIR3DL2*, they share a partial proximal motif with *KIR3DL1* and a full motif with *KIR3DL2*. The motif ambiguity between the *KIR3DL1/KIR3DL2* fusion and *KIR3DL2* can be resolved by linkage disequilibrium with *KIR2DS4*, since the fusion lacks *KIR2DS4* and the haplotypes with *KIR2DS4* cannot contain the fusion. Both systems consider the cB05 *KIR2DS2/KIR2DS3* fusion to be *KIR2DS2* (KIR2DS2*005 in IPD-KIR).

A total of 27 new haplotypes were added to the human genome reference as part of this study (accessions MN167504-30), 24 of which are from African Americans. The haplotypes cluster by structure, not population. Haplotypes from Africans or African Americans now constitute 47% of the KIR alternate references in the human genome project, and the human genome project contains more than three times as many alternative references for KIR than any full chromosome. These new haplotypes contain the first deposits of unusual linkages such as KIR2DL5B*00804 in the *KIR2DL5A* locus (MN167506), *KIR3DL1* or *KIR3DL2* (fusion) without *KIR2DS4* (tA02), *KIR2DL2* without *KIR2DS2* (cA03), and a *KIR2DS2* (fusion) without *KIR2DL2* (cB05). MN167526's *KIR2DS2* allele has a one base deletion after 285th base in the CDS sequence, which leads to a premature stop code at position 435 in the CDS.

Scisco Genetics' contribution to the human genome project of the KIR haplotypes has—like the human genome project itself—been an important advancement and will support the downstream discovery of the human immune system. Building on that work, the motifs and alignments presented here provide a means to help unify interpretation of the entire KIR region. They can be used to precisely define KIR haplotypes and loci, provide context for assigning alleles (especially fusion alleles) to genes, improve evolutionary inferences, improve imputation, interpret co-expression, and generate markers. The motif probes have been applied to a workflow to capture, assemble, and annotate KIR haplotypes at https://github.com/droeatumn/kass; it includes the ability to annotate KIR contigs/haplotypes as a separate workflow. The DNA alignments have been applied to discover markers used in a workflow to genotype KIR presence/absence from WGS at https://github.com/droeatumn/kpi.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by National Marrow Donor Program. The patients/participants provided their written informed consent to participate in this study. No animal studies are presented in this manuscript. No potentially identifiable human images or data is presented in this study.

## AUTHOR CONTRIBUTIONS

DR and RK designed the experiments. DR performed the *in silico* experiments and wrote the majority of the manuscript. C-WP and DG carried out KIR haplotype sequencing. All authors contributed to the article and approved the submitted version.

## FUNDING

Supported by grants N00014-18-2888 and N00014-20-1-2705 from the Department of the Navy, Office of Naval Research. The CIBMTR is supported primarily by Public Health Service U24CA076518 from the National Cancer Institute (NCI), the National Heart, Lung and Blood Institute (NHLBI) and the National Institute of Allergy and Infectious Diseases (NIAID); U24HL138660 from NHLBI and NCI; OT3HL147741, R21HL140314, and U01HL128568 from the NHLBI; HHSH250201700006C, SC1MC31881-01-00, and HHSH250201700007C from the Health Resources and Services Administration (HRSA); and N00014-18-1-2850, N00014-18-1-2888, and N00014-20-1-2705 from the Office of Naval Research. Additional federal support is provided by P01CA111412, R01CA152108, R01CA215134, R01CA218285, R01CA231141, R01AI128775, R01HL129472, R01HL130388, R01HL131731, U01AI069197, U01AI126612, and BARDA. The views expressed in this article do not reflect the official policy or position of the Department of the Navy, the Department of Defense or any other agency of the U.S. Government.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.585731/full#supplementary-material

**SUPPLEMENTARY TABLE 1 |** Motif assignments. The first column lists the alignment order of a pair of probes. The second column assigns a character to the pair.

**SUPPLEMENTARY TABLE 2 |** Allelic haplotypes. Each row represents a haplotype and columns D-X contain the presence/absence of the gene and allele names from IPD-KIR 2.9.0. Column A contains and ID consisting of the haplotype structure and GenBank accession. Column B contains the population of the individual. Column C contains the GenBank accession of the haplotype.

**SUPPLEMENTARY FIGURE 1 |** Phylogenetic tree made from the phylogenetic tree of the probe motifs of 68 human haplotypes and 1 chimpanzee in **Supplemental Data Sheet 2**. The relatively short chimpanzee haplotype is an outlier on the bottom of the tree and the relatively long cA01~tB04 is an outlier on the top.

**SUPPLEMENTARY FIGURE 2 |** Jalview overview of the MAFFT-generated multiple sequence alignment of the DNA of 68 human haplotypes plus 1 chimpanzee (**Supplementary Data Sheet 4**).

**SUPPLEMENTARY DATA SHEET 1 |** Zipped fasta file of the multiple sequence alignment of the probe motifs of 67 human haplotypes. Human haplotype cA01~tB04 is not included for purposes of visualization in the manuscript.

**SUPPLEMENTARY DATA SHEET 2 |** Zipped fasta file of the multiple sequence alignment of the probe motifs of 68 human haplotypes and 1 chimpanzee. The phylogenetic tree is in **Supplementary Image 1**.

**SUPPLEMENTARY DATA SHEET 3 |** Multiple sequence alignment of the DNA of 67 human haplotypes. Human haplotype cA01~tB04 is not included for purposes of visualization in the manuscript.

**SUPPLEMENTARY DATA SHEET 4 |** MAFFT-generated multiple sequence alignment of the DNA of 68 human haplotypes plus 1 chimpanzee. The overview image is in **Supplementary Image 2**.

## REFERENCES

1. Wroblewski EE, Parham P, Guethlein LA. Two to Tango: Co-evolution of Hominid Natural Killer Cell Receptors and MHC. *Front Immunol* (2019) 10:177. doi: 10.3389/fimmu.2019.00177
2. Bastidas-Legarda LY, Khakoo SI. Conserved and variable natural killer cell receptors: diverse approaches to viral infections. *Immunology* (2019) 156:319–28. doi: 10.1111/imm.13039
3. Díaz-Peña R, de los Santos MJ, Lucia A, Castro-Santos P. Understanding the role of killer cell immunoglobulin-like receptors in pregnancy complications. *J Assist Reprod Genet* (2019) 36:827–35. doi: 10.1007/s10815-019-01426-9
4. Lau CM, Adams NM, Geary CD, Weizman O-E, Rapp M, Pritykin Y, et al. Epigenetic control of innate and adaptive immune memory. *Nat Immunol* (2018) 19:963–72. doi: 10.1038/s41590-018-0176-1
5. Parham P, Guethlein LA. Genetics of Natural Killer Cells in Human Health, Disease, and Survival. *Annu Rev Immunol* (2018) 36:519–48. doi: 10.1146/annurev-immunol-042617-053149
6. Bruijnesteijn J, van der Wiel MKH, de Groot N, Otting N, de Vos-Rouweler AJM, Lardy NM, et al. Extensive Alternative Splicing of KIR Transcripts. *Front Immunol* (2018) 9:2846. doi: 10.3389/fimmu.2018.02846
7. Pende D, Falco M, Vitale M, Cantoni C, Vitale C, Munari E, et al. Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. *Front Immunol* (2019) 10:1179. doi: 10.3389/fimmu.2019.01179
8. Agrawal S, Prakash S. Significance of KIR like natural killer cell receptors in autoimmune disorders. *Clin Immunol* (2020) 216:108449. doi: 10.1016/j.clim.2020.108449
9. Chauhan SKS, Koehl U, Kloess S. Harnessing NK Cell Checkpoint-Modulating Immunotherapies. *Cancers* (2020) 12:1807. doi: 10.3390/cancers12071807
10. Marsh SGE, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics* (2003) 55:220–6. doi: 10.1007/s00251-003-0571-z
11. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* (2014) 43(D1):D423–431. doi: 10.1093/nar/gku1161
12. Pyo C-W, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different Patterns of Evolution in the Centromeric and Telomeric Regions of Group A and B Haplotypes of the Human Killer Cell Ig-Like Receptor Locus. *PLoS One* (2010) 5:e15115. doi: 10.1371/journal.pone.0015115
13. Pyo C-W, Wang R, Vu Q, Cereb N, Yang SY, Duh F-M, et al. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics* (2013) 14:89. doi: 10.1186/1471-2164-14-89
14. Roe D, Williams J, Ivery K, Brouckaert J, Downey N, Locklear C, et al. Efficient Sequencing, Assembly, and Annotation of Human KIR Haplotypes. *bioRxiv* (2020). doi: 10.1101/2020.07.12.199570
15. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, et al. Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc Natl Acad Sci* (2000) 97:4778–83. doi: 10.1073/pnas.080588597

16. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun* (2017) 18:127–34. doi: 10.1038/gene.2017.10
17. Carrillo-Bustamante P, Keşmir C, de Boer RJ. The evolution of natural killer cell receptors. *Immunogenetics* (2016) 68:3–18. doi: 10.1007/s00251-015-0869-7
18. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* (2019) 34:1155–62. doi: 10.1038/s41587-019-0217-9
19. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings Bioinf* (2017) 20(4):1160–6. doi: 10.1093/bib/bbx108
20. Falco M, Moretta L, Moretta A, Bottino C. KIR and KIR ligand polymorphism: a new area for clinical applications?: Relevance of KIR and KIR-L polymorphism. *Tissue Antigens* (2013) 82:363–73. doi: 10.1111/tan.12262
21. Ordóñez D, Meenagh A, Gómez-Lozano N, Castaño J, Middleton D, Vilches C. Duplication, mutation and recombination of the human orphan gene KIR2DS3 contribute to the diversity of KIR haplotypes. *Genes Immun* (2008) 9:431–7. doi: 10.1038/gene.2008.34
22. Cisneros E, Moraru M, Gómez-Lozano N, López-Botet M, Vilches C. KIR2DL5: An Orphan Inhibitory Receptor Displaying Complex Patterns of Polymorphism and Expression. *Front Immunol* (2012) 3:289. doi: 10.3389/fimmu.2012.00289

**Conflict of Interest:** C-WP and DG are employees of Scisco Genetics.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.4

ORIGINAL RESEARCH
published: 26 November 2020
doi: 10.3389/fimmu.2020.583013

# Accurate and Efficient KIR Gene and Haplotype Inference From Genome Sequencing Reads With Novel K-mer Signatures

*David Roe*[1]* *and Rui Kuang*[1,2]

*[1] Bioinformatics and Computational Biology, University of Minnesota, Rochester, MN, United States, [2] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, United States*

The killer-cell immunoglobulin-like receptor (KIR) proteins evolve to fight viruses and mediate the body's reaction to pregnancy. These roles provide selection pressure for variation at both the structural/haplotype and base/allele levels. At the same time, the genes have evolved relatively recently by tandem duplication and therefore exhibit very high sequence similarity over thousands of bases. These variation-homology patterns make it impossible to interpret KIR haplotypes from abundant short-read genome sequencing data at population scale using existing methods. Here, we developed an efficient computational approach for *in silico* KIR probe interpretation (KPI) to accurately interpret individual's KIR genes and haplotype-pairs from KIR sequencing reads. We designed synthetic 25-base sequence probes by analyzing previously reported haplotype sequences, and we developed a bioinformatics pipeline to interpret the probes in the context of 16 KIR genes and 16 haplotype structures. We demonstrated its accuracy on a synthetic data set as well as a real whole genome sequences from 748 individuals from The Genome of the Netherlands (GoNL). The GoNL predictions were compared with predictions from SNP-based predictions. Our results show 100% accuracy rate for the synthetic tests and a 99.6% family-consistency rate in the GoNL tests. Agreement with the SNP-based calls on KIR genes ranges from 72%–100% with a mean of 92%; most differences occur in genes *KIR2DS2*, *KIR2DL2*, *KIR2DS3*, and *KIR2DL5* where KPI predicts presence and the SNP-based interpretation predicts absence. Overall, the evidence suggests that KPI's accuracy is 97% or greater for both KIR gene and haplotype-pair predictions, and the presence/absence genotyping leads to ambiguous haplotype-pair predictions with 16 reference KIR haplotype structures. KPI is free, open, and easily executable as a Nextflow workflow supported by a Docker environment at https://github.com/droeatumn/kpi.

**Keywords: killer-cell immunoglobulin-like receptor, genotype, haplotype, interpretation, natural killer, whole genome sequencing (WGS)**

OPEN ACCESS

**Edited by:**
*Malini Raghavan,*
*University of Michigan, United States*

**Reviewed by:**
*Sandeep Kumar Dhanda,*
*St. Jude Children's Research Hospital,*
*United States*
*Marcia Maria Marques,*
*Barretos Cancer Hospital, Brazil*

***Correspondence:**
*David Roe*
*roex0050@umn.edu*

**Specialty section:**
*This article was submitted to*
*Alloimmunity and Transplantation,*
*a section of the journal*
*Frontiers in Immunology*

**Received:** *13 July 2020*
**Accepted:** *02 November 2020*
**Published:** *26 November 2020*

**Citation:**
*Roe D and Kuang R (2020) Accurate and Efficient KIR Gene and Haplotype Inference From Genome Sequencing Reads With Novel K-mer Signatures.*
*Front. Immunol. 11:583013.*
*doi: 10.3389/fimmu.2020.583013*

## INTRODUCTION

Human chromosome 19q13.4 contains a ~150–250 kilobase region encoding 16 genes of the natural killer-cell immunoglobulin-like receptor (KIR) family. These genes are ~4–16 kilobases long and evolved *via* tandem duplication during primate evolution (1, 2). The KIR receptors recognize human leukocyte antigen (HLA) class I molecules and contribute to natural killer (NK) cell functions *via* activating or inhibiting signals. These receptor-ligand pairs coevolved under selection pressures from reproduction and pathogenic defense (3), and it is believed that KIR genes have undergone a balancing selection *via* duplications and deletions into two broad categories of haplotypes, in which one category tends to vary more at the allelic level and the other tends to vary more at the structural (gene content and order) level (4–6). A few dozen KIR full haplotype sequences and approximately 2500 full- or inter-gene sequences have been publicly deposited (5, 7, 8). Haplotype structures are divided into two classes (9). Class 'A' contains one haplotype and its deleted forms. Class 'B' haplotypes are more structurally diverse and contain a variety of insertions and deletions. Generally, the A haplotype occurs with 50-60% frequency, haplotypes that are half-A and half-B occur with 30-40%, and the rest of the haplotypes are variants of the B haplotypes. Except for some rarer deleted forms, KIR haplotypes are structurally variable around 4 'framework' genes (*KIR3DL3*, *KIR3DP1*, *KIR2DL4*, *KIR3DL2*), with *KIR3DL3* through *KIR3DP1* defining the proximal (or 'centromeric') region and *KIR2DL4* through *KIR3DL2* defining the distal (or 'telomeric') region, with the two gene-rich regions separated by the relatively large and recombinant *KIR3DP1-KIR2DL4* intergene region.

It is difficult to interpret the KIR region with high-throughput sequencing reads for an individual human genome when the structural arrangements are unknown; indeed, it is difficult even when the structural haplotypes are known, since the read length is too short to map unambiguously to the repetitive and homologous KIR genes. As a consequence, the reads from KIR region are ignored, as to the best of our knowledge, there are currently no algorithms to interpret KIR from whole genome sequencing (WGS). SNP (single nucleotide polymorphism)-based KIR interpretation is more commonly applied. For example, KIR*IMP is a web-application to predict genes and haplotypes from microarray SNP genotypes (10). As an algorithm whose raw data is microarray calls, KIR*IMP can interpret KIR from genome wide SNP arrays, but it is not applicable to interpret KIR from raw sequences.

Since a general solution for KIR structural interpretation from raw genomic DNA is not currently available, this study implements such an algorithm for the prediction of KIR genes and full structural haplotypes from any type of raw full-region-or-greater genomic sequence at population scale. In particular, we systematically evaluated small markers for KIR genes and then applied those markers to a synthetic KIR probe interpretation (KPI) algorithm for the presence/absence of 16 KIR genes and 16 haplotype structures. Our approach leverages recent bioinformatics innovations for short sequence ('probe') genotyping, along recently published KIR reference haplotypes. The KPI algorithm first efficiently counts the occurrence of each kmer probe in the raw sequences, and then uses multiple probes per gene to call its presence/absence. Those 16 genotypes are then used to generate haplotype-pair predictions. In the experiments, we report 100% accuracy on a test set of synthetic haplotypes for comparisons with known truth. We also report that gene and haplotype-pair predictions for the WGS GoNL cohort are family consistent and compare favorably with reference frequencies in comparison to SNP-based predictions using KIR*IMP.

## MATERIALS AND METHODS

### Overview

The workflow of KPI consists of three steps,

1. Discover the 25mer gene markers based on a multiple sequence alignment analysis of 68 full-length haplotype sequences.
2. Count the 25mer markers in the reads of genomic DNA per individual to generate the individual's 25mer genotype.
3. Predict presence/absence per gene from the marker genotypes for each individual.
4. Predict haplotype pairs from the gene presence/absence calls for each individual.

In the following, we first explain each step and then describe the synthetic data and GoNL data used for the evaluation.

### Step 1: Discovering 25mer Gene Markers

To discover gene marker 25mers, first a multiple sequence alignment (MSA) was created with 68 publicly deposited full-length haplotypes sequences (11). Briefly, each haplotype was annotated at an average resolution of ~4kbp using a set of 15 120-base markers. This high-level annotation was aligned into a MSA representing a structural alignment of all haplotypes. Then, each subregion was aligned to base pair resolution. This resulted in a full resolution, full haplotype MSA that accurately classifies each allele into a haplotype-defined locus, and it aligns the alleles precisely at each locus. The haplotype and gene annotations of the MSA provided a list of full-length alleles for 16 genes: *KIR2DL1-5, KIR2DS1-5, KIR2DP1, KIR3DL1-3, KIR3DP1*, and *KIR3DS1*. Markers for each gene locus were chosen by selecting all sequences of length 25 (25mers) present in every allele of the gene but not elsewhere in the KIR haplotypes nor the rest of the genome reference GRCh38. More details about the algorithm are in **Supplemental Figure 1**. The marker sequences are in **Supplemental Data Sheet 1** and also checked in to GitHub at https://github.com/droeatumn/kpi/tree/master/input in text and fasta format.

### Step 2: Count 25mer Markers

KMC 3, with workflows implemented in Nextflow (12) and Apache Groovy (13) and a software environment implemented as a Docker container, is used to create 25mer databases from sequence or short-read data and match the markers across the datasets. Using KMC 3, we generate the list of all 25mers from the short reads of each individual and then match the 25mers in the marker databases to report the hit counts of each 25mer marker in the individual. Details are in **Supplemental Figure 1**.

## Step 3: Individual Genotyping From 25mer Markers

KPI calls presence/absence per gene by aggregating the presence/absence genotypes of many small (25mer) markers, each specific to one gene. 25mers with hit counts less than three are considered sequencing errors and set to zero. If the mean hit count of all the markers per gene is zero, then the gene is predicted absent; otherwise, it is called present. Additional details can be found in **Supplemental Figure 1**.

## Step 4: Individual Haplotyping From Genotypes

Haplotype-pair predictions were made by fitting the genotype to all possible pairs of the 16 structural reference haplotypes defined in **Figure 1**. The numbers and frequencies of the haplotypes are from Jiang et al. 2012 (4) (**Table 1**); some of their haplotypes are combined because Jiang et al. consider certain alleles as separate haplotypes, such as full or deleted alleles of *KIR2DS4*. These 16 haplotypes represent 97% of all haplotypes in the Jiang et al. report.

For the GoNL predictions, haplotype ambiguity was reduced by family trio patterns and then further by the EM (Expectation-Maximization)-based methods as described and used in Vierra-Green 2012 (14). Haplotype frequencies were calculated from the EM-reduced individual haplotype-pair predictions. These haplotype frequency calculations are not possible on the KPI's haplotype-pair predictions because they can be ambiguous.

## Synthetic Capture on Diploid Data

KPI was evaluated on a synthetic test set. There are six reference haplotype structures with publicly deposited full-length sequences (**Figure 1**, top six rows). For each of these six structures, one sequence was randomly chosen to represent that structure, and it was paired with a random haplotype sequence from the set of all sequences, with an equal probability for each sequence. dwgsim (12) was used to generate 10,000 2×150 pair reads per haplotype (~20× coverage) with 1% error rate. This provided a simulated six-person validation set of six diploid whole-region short-read sequences, representing all fully sequenced haplotype structures and paired to provide a variety of genotypes. The sequences are included in **Supplemental Data Sheet 2**.

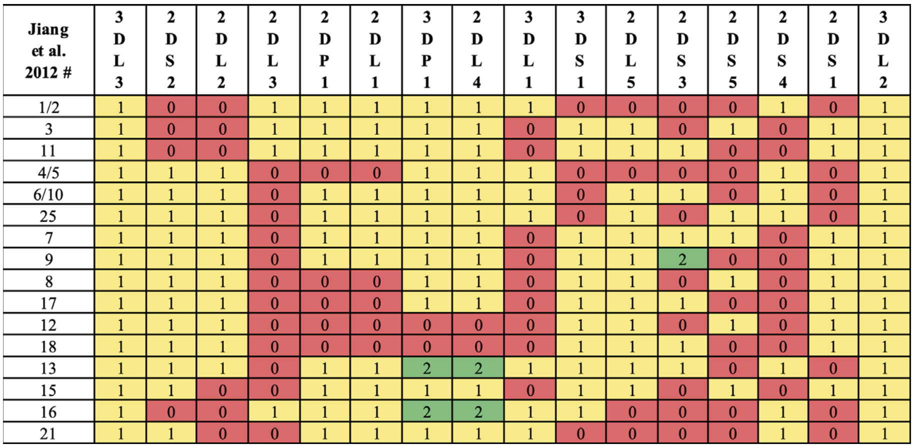
**TABLE 1** | Reference haplotype names and frequencies.

| Jiang et al. 2012 # | informal names | Jiang et al. 2012 freq. |
|---|---|---|
| 1/2 | cA01~tA01 | 55.2% |
| 3 | cA01~tB01_2DS5 | 10.9% |
| 11 | cA01~tB01_2DS3 | 1.4% |
| 4/5 | cB02~tA01 | 12.8% |
| 6/10 | cB01~tA01_2DS3 | 6.9% |
| 25 | cB01~tA01_2DS5 | 0.1% |
| 7 | cB01~tB01_2DS3_2DS5 | 2.6% |
| 9 | cB01~tB01_2DS3_2DS3 | 2.1% |
| 8 | cB02~tB01_2DS5 | 2.1% |
| 17 | cB02~tB01_2DS3 | 0.3% |
| 12 | cB04~tB03_2DS5 | 0.8% |
| 18 | cB04~tB03_2DS3 | 0.3% |
| 13 | cB01~tB05 | 0.7% |
| 15 | cB05~tB01 | 0.4% |
| 16 | cA01~tB05 | 0.3% |
| 21 | cB05~tA01 | 0.2% |
| sum | | 97.0% |

*The first column contains the numeric label assigned to haplotypes in Jiang et al. (2012). Column 2 contains the informal names along with a reference frequency in column 3.*

## GoNL Family WGS and Immunochip SNP Data

KPI was also run on a large real-world example. WGS was obtained from The Genome of the Netherlands (GoNL) (13), a



| Jiang et al. 2012 # | 3DL3 | 2DS2 | 2DL2 | 2DL3 | 2DP1 | 2DL1 | 3DP1 | 2DL4 | 3DL1 | 3DS1 | 2DL5 | 2DS3 | 2DS5 | 2DS4 | 2DS1 | 3DL2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1/2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 3 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 11 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 4/5 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 6/10 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| 25 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 7 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 9 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 1 |
| 8 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 17 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 12 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 18 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 13 | 1 | 1 | 1 | 0 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| 15 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 16 | 1 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 21 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

**FIGURE 1** | Reference haplotype definitions. Haplotype numeric labels (Jiang et al. 2012) are shown with their definition *via* gene counts. Following Jiang et al. convention, some haplotypes (e.g., 7, 9) are distinguished by *KIR2DS3/KIR2DS5* alleles instead of structural differences. In this study, some haplotypes (e.g., 1, 2) are combined, as *KIR2DS4* full/deleted alleles are not considered in KPI's genotyping.

genome sequencing project whose goal is to map the genetic variation within the population of Netherlands in 250 family trios (750 individuals). The project provided non-paired sequencing of the whole genomes of the population, which was done on the Illumina HiSeq 2000 platform. Coverage of the KIR region were similar to the previously reported (13) whole-genome average of ~10–15×. Two individuals from two different families were removed from the GoNL project for data quality reasons, giving a total of 748 individuals.

KPI's GoNL predictions were compared with results from microarray-based interpretation algorithm KIR*IMP. Illumina Immunochip microarray SNP data was obtained from GoNL (13). The data was prepared and executed following instructions using KIR*IMP v1.2.0 on 2019-10-05.

To the best of our knowledge, there only exists one method to predict KIR gene content from WGS sequences (15). However, we were unable to obtain results with it for both evaluation data sets. According to the authors, the current version is deprecated and to be replaced soon (16).

# RESULTS

The predictions were evaluated in the small synthetic test set, where truth is known and a large real-world test set, where truth is unknown except for family relationships. Predictions were evaluated by comparing gene and haplotype-pair predictions to: known truth in the synthetic cohort, and family consistency (real-world cohort only), reference frequencies from Jiang et al.'s family copy number study (4), and the Allele Frequencies Database (17) in the real-world cohort. The real-world cohort was also compared with predictions from microarray-based algorithm KIR*IMP, although KIR*IMP was not considered ground truth as it reports accuracies as low as 81% for some genes (10). Haplotype-pair predictions were considered to be family consistent if each parents' two haplotype predictions contained at least one of the child's two predictions and one of the child's haplotypes occurred in one parent and the other haplotype occurred in the other parent.

## Synthetic Evaluation

**Table 2** shows the results of the synthetic tests. The gene present/absent calls were 100% accurate for all genes. Although the haplotype predictions are ambiguous in half of the individuals, all are consistent with the ground truth.

## GoNL Evaluation

**Table 3** shows a summary of the gene prediction results from KPI and KIR*IMP on the GoNL data set. A reference frequency range is included from Allele Frequencies Net, selecting European cohorts >= 500 individuals. Overall agreement between KIR*IMP and KPI for the 16 genes (**Table 3**, column 6) ranges from 72% to 100%, with a mean of 92%. KIR*IMP differs from the reference haplotype (**Table 3**, column 7) frequency range by >10% in four genes (*KIR2DS2*, *KIR2DL2*, *KIR2DL5*, and *KIR2DS3*) compared with 0 genes for KPI (**Table 3**, column 8). Both KIR*IMP and KPI differ from the *KIR2DS1* reference by 9-10%, although the two algorithms agree in 98% of individuals for that gene.

**Table 4** breaks down the differences between KIR*IMP and KPI in a confusion matrix. In the cases where KIR*IMP calls present ('P') and KPI calls absent ('A') (**Table 4**, column 2), the largest discrepancies are found in the centromeric genes *KIR2DS2* (8%), *KIR2DL2* (8%), and *KIR2DL3* (6%). In the reverse cases, when KIR*IMP calls absent and KPI calls present (**Table 4**, column 3), the largest discrepancies are greater and occur with the centromeric *KIR2DS2* (20%), *KIR2DL2* (20%), and the paralogous (centromeric or telomeric) *KIR2DL5* (11%), and *KIR2DS3* (19%).

**TABLE 2** | Results of synthetic tests.

| haplotype 1 structure | haplotype 2 structure | haplotype 1 GenBank accession | haplotype 2 GenBank accession | KPI haplotype prediction | haplotypes consistent w/ truth? | gene prediction accuracy |
|---|---|---|---|---|---|---|
| cA01~tA01 | cA01~tA01 | GU182344 | GU182340 | cA01~tA01 +cA01~tA01 | Y | 100% |
| cA01~tB01 | cA01~tA01 | KU645197 | GU182360 | cA01~tA01 +cA01~tB01 or cA01~tB01 +cA01~tB05 | Y | 100% |
| cB01~tA01 | cA01~tA01 | GU182351 | NC000019.10 | cA01~tA01 +cB01~tA01 | Y | 100% |
| cB01~tB01 | cB02~tA01 | GU182339 | GU182353 | 10 possibilities, including cB01~tB01 +cB02~tA01 | Y | 100% |
| cB02~tA01 | cA01~tA01 | GU182341 | KP420442 | cA01~tA01 +cB02~tA01 | Y | 100% |
| cB02~tB01 | cA01~tA01 | GU182359 | KP420439 | 9 possibilities, including cA01~tA01 +cB02~tB01 | Y | 100% |

*The first four columns detail the sequences from which the tests (n=6 haplotype-pairs) were generated. The fifth column is killer-cell immunoglobulin-like receptor probe interpretation's (KPI's) haplotype predictions, some of which are summarized for display.*

**TABLE 3 |** Summary of killer-cell immunoglobulin-like receptor (KIR)*IMP and KIR probe interpretation (KPI) gene predictions.

| gene | reference freq. | KIR*IMP freq. | KPI freq. | KIR*IMP - KPI | KIR*IMP & KPI agreement | KIR*IMP - reference | KPI - reference |
|---|---|---|---|---|---|---|---|
| 2DS2 | 53-54% | 39% | 52% | –13% | **72%** | **–14%** | **–1%** |
| 2DL2 | 53-54% | 39% | 51% | –12% | **72%** | **–14%** | **–2%** |
| 2DL3 | 90% | 96% | 92% | 4% | 92% | 6% | 2% |
| 2DP1 | 96% | 99% | 97% | 2% | 97% | 3% | 1% |
| 2DL1 | 96% | 99% | 97% | 2% | 97% | 3% | 1% |
| 3DP1 | 100% | 100% | 100% | 0% | 100% | 0% | 0% |
| 2DL4 | 100% | 100% | 100% | 0% | 100% | 0% | 0% |
| 3DL1 | 93%–94% | 96% | 96% | 0% | 100% | 2% | 2% |
| 3DS1 | 38%–44% | 33% | 35% | –2% | 97% | –5% | –3% |
| 2DL5 | 53%–56% | 38% | 47% | –9% | **87%** | **–15%** | **–6%** |
| 2DS3 | 30%–31% | 10% | 29% | –18% | **81%** | **–20%** | **–1%** |
| 2DS5 | 30%–36% | 25% | 27% | –2% | 96% | –5% | –3% |
| 2DS4 | 92%–94% | 96% | 96% | 0% | 100% | 2% | 2% |
| 2DS1 | 43%–44% | 33% | 34% | –1% | 98% | –10% | –9% |
| average | | | | | 92% | | |

*Frequencies relative to Genome of the Netherlands (GoNL) cohort of 748 individuals. The abbreviated gene name is in column 1. Column 2 lists the reference frequencies from The Allele Frequency Net Database. The predicted frequencies from KIR*IMP and KPI are in columns 3 and 4, respectively. The delta between KIR*IMP and KPI is shown in the column 5. Column 6 shows the agreement between KIR*IMP and KPI. Column 7 shows the delta between KIR*IMP and the reference. Column 8 shows the delta between KPI and the reference. Frequencies with differences >10% are in bold.*

**TABLE 4 |** Confusion matrix of killer-cell immunoglobulin-like receptor (KIR)*IMP and KIR probe interpretation (KPI) gene predictions.

| gene | KIR*IMP:P KPI:A | KIR*IMP:A KPI:P | KIR*IMP:P KPI:P | KIR*IMP:A KPI:A |
|---|---|---|---|---|
| 2DS2 | 8% | **20%** | 32% | 40% |
| 2DL2 | 8% | **20%** | 31% | 41% |
| 2DL3 | 6% | 2% | 90% | 2% |
| 2DP1 | 2% | 0% | 97% | 0% |
| 2DL1 | 2% | 1% | 97% | 0% |
| 3DP1 | 0% | 0% | 100% | 0% |
| 2DL4 | 0% | 0% | 100% | 0% |
| 3DL1 | 0% | 0% | 96% | 4% |
| 3DS1 | 1% | 3% | 32% | 64% |
| 2DL5 | 2% | **11%** | 36% | 51% |
| 2DS3 | 1% | **19%** | 10% | 71% |
| 2DS5 | 1% | 3% | 24% | 72% |
| 2DS4 | 0% | 0% | 96% | 4% |
| 2DS1 | 1% | 1% | 33% | 66% |

*Frequencies relative to GoNL cohort size of 748 individuals. The abbreviated gene names are in column 1. Column 2 lists the cases when KIR*IMP calls present ('P') and KPI calls absent ('A'). Column 3 lists the cases when KIR*IMP calls absent ('A') and KPI calls present ('P'). Column 4 is when they both call present. Column 5 is when they both call absent. Discrepancies >10% are in bold.*

Per-individual haplotype-pair predictions for the GoNL cohort are included in **Supplemental Table 1**. Three lists of haplotype-pairs are provided: one for the initial fitting of all possible haplotype-pairs that could explain the genotype (i.e., KPI's output); another that reduces those possibilities by family relationships; and one for the EM-reduced final haplotype-pair predictions.

KIR*IMP makes one most-likely prediction for all individuals. KPI's predictions are sometimes ambiguous, with most predictions (mode) having one haplotype-pair but a mean of 2.3, standard deviation of 2.5, and a maximum of 14 haplotype-pair predictions per individual in the context of the 16 reference haplotypes. The KIR*IMP predictions are family consistent 100% of the time compared with 99.6% for KPI. However, the haplotype pair predictions between the two algorithms are concordant only 58% of the predictions.

**Table 5** compares the haplotype predictions between KIR*IMP and the EM-reduced haplotype pair predictions from the KPI output. KIR*IMP fit 100% of its predicted genotypes into 15 of its reference haplotypes. KPI fit 97% of its predicted genotypes into its 16 reference haplotypes. KIR*IMP made predictions for two haplotypes (cA01~tB04 and cB04~tB03, numbered 14, 18, and 12), totaling 0.47%, that are not in KPI's set of reference haplotypes. KPI's haplotype-pair predictions are too ambiguous to summarize in haplotype frequencies.

## DISCUSSION

KPI was evaluated by Chen et al. as part of a larger effort (18). In a cohort of 72 individuals with ground truth determined by LinkSeq qPCR, Chen et al. report six mismatches in one sample (possibly swapped), and apart from this 95.8% accuracy for *KIR2DS3* and 100% accuracy for the 15 other genes. As they note, it is now possible to interpret HLA binding alleles and the presence/absence of all their KIR receptors from short-read high-throughput sequencing, and this combination is a valuable advancement for research and medicine. Indeed, they compare KPI favorably with respect to clinical accreditation standards.

The findings by Chen et al. are consistent with the results of our synthetic test, whose accuracy was 100% for all genes. One drawback of the design of the synthetic test is that the haplotypes used in the test were also included in the MSA that was used to generate the per-gene probes. However, the main purposes of the synthetic tests were to test the application of the markers to short reads in a variety of genotypes and recover their original geno- and haplo-types; this is value-added compared with simply demonstrating sequences unique to a gene. The almost-perfect results of the Genentech and synthetic experiments, along with a GoNL results that had a 99.6% family-consistency rate and in line with expected frequencies, provide evidence that KPI's gene predictions are very accurate.

The evidence also suggests that KPI's haplotype results are accurate, although often ambiguous: the accuracy in the

**TABLE 5** | Comparison of killer-cell immunoglobulin-like receptor (KIR)*IMP highest probability and EM-reduced haplotype prediction frequencies.

| hap # | reference frequency | KIR*IMP | KPI w/ EM | KIR*IMP - reference | KPI w/ EM - reference | KIR*IMP - KPI w/ EM |
|---|---|---|---|---|---|---|
| 1 | 55.20% | 71.86% | 59.70% | 16.70% | 4.50% | 12.17% |
| 2 | | | | | | |
| 3 | 10.90% | 12.57% | 9.60% | 1.70% | –1.29% | 2.95% |
| 11 | 1.40% | 1.47% | 0.60% | 0.00% | –0.82% | 0.85% |
| 4 | 12.80% | 7.49% | 15.30% | –5.30% | 2.55% | -7.83% |
| 5 | | | | | | |
| 9 | 2.10% | 3.41% | 2.90% | 1.30% | 0.78% | 0.52% |
| 7 | 2.60% | 0.33% | 3.60% | –2.20% | 1.00% | –3.24% |
| 6 | 6.90% | 1.80% | 5.90% | –5.10% | –1.08% | –4.05% |
| 10 | | | | | | |
| 8 | 2.10% | 0.60% | 0.50% | –1.50% | –1.65% | 0.12% |
| 17 | 0.30% | 0.00% | 0.00% | –0.30% | –0.23% | –0.03% |
| 14* | 2.40% | 0.40% | 0.00% | –2.00% | 0.00% | 0.00% |
| 18* | 0.30% | 0.07% | 0.00% | –0.20% | 0.00% | 0.00% |
| 12* | 0.80% | 0.00% | 0.00% | –0.80% | 0.00% | 0.00% |
| mean | 97.00% | 100.00% | 98.10% | | | |

*The table shows the comparison of the predictions between both methods as well as with reference European frequencies from Jiang et al. 2012 (column 2), which is the source of the haplotype numbers (column 1). KIR*IMP's haplotype frequencies for the 1496 GoNL haplotypes are in column 3; some haplotypes are combined, as the haplotype numbers distinguish KIR2DS4 alleles. Column 4 contains frequencies for EM-reduced KIR probe interpretation (KPI) haplotype predictions Column 5 compares KIR*IMP frequencies with the reference, as column 6 does for EM predictions. Finally, column 7 compares the frequencies of KIR*IMP and the EM-reduced predictions. Haplotypes with a predicted frequency of 0 in both KIR*IMP and KPI are not shown. Haplotypes 14, 18, and 12 are in KIR*IMP's set of reference haplotypes, but not KPI's.*

synthetic test was 100% and the GoNL family-consistency was 99.6, and the predictions allow EM predictions that align with the expected population frequency from the literature.

KIR*IMP's haplotype frequency estimations differ from expectations in some areas. The evidence from comparisons with frequency reports from Jiang et al. 2012 (**Table 5**, column 5) suggest KIR*IMP overestimated cA01~tA01 (haplotype numbers 1 and 2) and underreported haplotypes containing cB01 or cB02 centromeric regions combined with the tA01 telomeric region (cB01~tA01 and cB02~tA01) in the GoNL cohort. This discrepancy can also be seen in the predicted genotype frequencies, where KIR*IMP relatively under calls the presence of *KIR2DS2*, *KIR2DL2*, and *KIR2DS3* by ~20% and *KIR2DL5* by ~10% compared with KPI and the historical European frequencies from Allele Frequency Net database (**Table 5**, column 6); all four of those genes are in cB01, and *KIR2DS2* and *KIR2DL2* are also in cB02. GoNL genotyping was done on the Immunochip, which is the best option according to the KIR*IMP manuscript. With that chip, they report accuracies of 100% for *KIR2DS2*, 98% for *KIR2DL2*, 82% for *KIR2DL5*, 81% for *KIR2DS3*, and 95% for *KIR2DS5* in their Norwegian-German validation cohort. Although the family consistency rate is 100% for KIR*IMP and 96.6% for KPI, their haplotype-pair predictions only agree in 58% of individuals. Without ground truth available, without any reason to expect this cohort to deviate from expectations, and considering KIR*IMP's self-reported accuracy, the evidence suggests that KPI's predictions are more accurate than KIR*IMP's in this cohort and specifically that KIR*IMP under called the presence of genes *KIR2DS2*, *KIR2DL2*, *KIR2DS3*, *KIR2DL5* and haplotypes cB01~tA01 and cB02~tA01. As reviewed recently by Wright et al., this may be particularly relevant in the context of hematopoietic stem cell transplantation, where some case/control studies claim an important role for these regions (19). There are several potential reasons KIR*IMP's predictions may be less accurate than KPI's. The reference haplotypes used for marker discovery for KIR*IMP were defined by copy number genotyping and family relationships; KPI defined its haplotypes using a MSA of full haplotype sequences. KIR*IMP's input is restricted to a few hundred single nucleotide polymorphisms, whereas KPI can use the entire genomic range of KIR sequences of length 25, which provides the potential for more information per marker and a broader base of markers. KIR*IMP uses a small number of SNPs to call one or more genes, whereas KPI uses dozens-to-thousands of 25mers to call a single gene, One of the steps of KIR*IMP's workflow is to align and phase all the SNPs to one 'A' haplotype, which may be a limitation for genes not on that haplotype; all the gene and haplotypes we found to have lower accuracy rates are not located on the 'A' haplotype. KPI has no alignment or assembly steps. It is also important to note that the primary purpose for the comparison with KIR*IMP was not to evaluate the potential success of predicting KIR genes and haplotypes using SNPs vs sequence reads, but rather to compare the two algorithms. Although both algorithms predict the presence/absence of KIR genes and structural haplotypes, their solution domains are very different: microarray SNP panels vs raw genomic DNA reads. Both algorithms report the lowest accuracy rates for *KIR2DS3* and a ~10% lower frequency rate for *KIR2DS1* in GoNL compared with reference frequencies.

The 85% family consistency rate of the EM-reduced haplotype predictions suggest that KIR haplotype ambiguity cannot be accurately reduced at the individual level *via* expectation-maximization. However, since the EM-reduced haplotype frequencies are in line with references, it is possible the predictions might aggregate to population-level in a maximum-likelihood manner and therefore perhaps may still be useful for some population genetics purposes.

Traditional lab-based SSO presence/absence genotyping relies on a single short-sequence strategy, an approach that can be applied similarly to synthetic analysis of large amounts of WGS. In this virtual context, primer locations are not needed, and kmer searching is efficient and accurate at populations scales. To develop this synthetic SSO-like (kmer) library, we leveraged the information from a multiple sequence alignment of all full-length haplotypes that are available for this study. We believe this is a more accurate approach than using IPD-KIR reference alleles, because the IPD-KIR reference alleles do not require the haplotype location to be known. In addition, fusion alleles are assigned in IPD-KIR to one of the two parent genes, and therefore large sequences of some alleles are not really from the gene in which they are classified. We used 25 for our 'k' (i.e., sequence size) because BLAST searching indicated this to be a conservative minimum length needed to distinguish a small set of test markers to the KIR region. We did not experiment with any k size other than 25, since the choice gives a reasonable number of significant markers and their lack of off-KIR hits as tested in the GoNL population WGS confirms the effectiveness as gene/intergenic markers. The only fundamental benefit to shorter markers would be in the case when there were no longer markers; however, 25mer markers were found for every gene. The only fundamental benefit to longer markers would be if the markers were not unique to the region; however, all the markers are unique to their region. Many of the 25mers overlap each other, effectively simulating a single longer marker. Similar to the reasoning about probe length, probe mismatches would only need to be relaxed if a locus did not have any markers. Since at least one marker was discovered for each gene, mismatches did not need to be incorporated. Having thus obtained the region markers, we then used the most common ('peak') hit count from each gene/intergene's library of sequences to make the PA genotype calls (**Supplemental Figure 1**). This adds a certain amount of allelic flexibility in the algorithm because the ultimate call is an average of all the markers for that gene; if some markers miscall, the overall call for the gene will be unaffected if the majority of the other markers are accurate. Since KPI decomposes the genetic information into 25mers, it works with any collection of DNA reads, as long as the KIR region is included. It works with fasta, fastq, single, paired, short, and long reads. Since the markers are not unique to exons, it will not work with cDNA or exon only reads.

One limitation of the method is that the markers do not mark DNA segments longer than one gene. Perhaps this is primarily due to the frequent recombination between haplotypes. Although recombination has been reported in multiple loci, the hotspot in between the centromeric and telomeric regions is particularly strong, and, in general, any pairing of the two can be expected. We evaluated single markers for haplotypes, but we did not find any. This is particularly relevant in light of the observation that haplotyping from genotypes seems to have limited accuracy under maximum likelihood assumptions (**Table 5**). It is possible that an algorithm that uses applies multiple markers in a hierarchical or combinational manner may be more successful. For future work, we plan to further evaluate *KIR2DS3* (95.8% accuracy in the Chen et al. evaluation) and evaluate the genotyping in diverse populations. It is possible that population robustness is a weakness of the method, although the fact that almost half of individuals in the discovery cohort are African or African American provides some optimism.

For the WGS data set, KPI averaged less than 1 h of computing time per individual, with 32 cores of CPU and 32G RAM. The majority of the time is spent using KMC 3 to build the kmer database. kmer counting is an active area of research. Since KPI can easily be altered to use any such application, it has potential for future efficiency improvements.

The markers discovered in this study were enabled by a full-haplotype MSA, as described recently by Roe et al. (11). That manuscript makes observations about the composition and order of sequences within KIR haplotypes, and it reports the implications for our understanding of the relationship between haplotypes, loci, and genes. Here, we have leveraged that basic understanding for practical use by developing a free and open interpretation application, evaluating a SNP interpretation algorithm, and contributing KIR interpretation to an important population genetics resource for Netherlands and many types of human genetic researchers. We described how the gene markers were discovered from the MSA, and we demonstrated their use to predict genes and haplotype-pairs at high accuracy (97%+) with population scale from any kind of sequence data that includes the full KIR region, including WGS. It was tested on synthetic ground-truth sequences and a large cohort of family WGS. In addition, we compared our algorithm to the leading SNP-based interpretation algorithm. KPI is free software with a GPL3 license and is implemented as a Nextflow workflow backed with an optional Docker environment. It is available at https://github.com/droeatumn/kpi.

## AUTHOR'S NOTE

This manuscript has been released as a pre-print at bioRxiv with the same title and authors (20).

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Minnesota. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

DR and RK designed the experiments and wrote the manuscript. DR conducted the experiments. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

This study makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI). The samples were provided by The LifeLines Cohort Study (21), and generation and management of GWAS genotype data for it, is supported by the Netherlands Organization of Scientific Research (NWO, grant 175.010.2007.006), the Dutch government's Economic Structure Enhancing Fund (FES), the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, the University Medical Center Groningen, the University of Groningen, the Dutch Kidney Foundation and Dutch Diabetes Research Foundation.

Thanks to Christian Hammer and his team from Genentech for assistance testing and debugging the application. Also, thanks to Cynthia Vierra-Green and Martin Maiers from the Center for International Blood and Marrow Transplant Research (CIMBTR) as well as Julia Udell from Mayo Clinic and the University of Minnesota for KIR consultation and consolation.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.583013/full#supplementary-material

**SUPPLEMENTAL FIGURE 1 |** Algorithm details. Data sheet 3 is a Microsoft Word document containing the commands to query the markers per genome and details as to how each region was genotyped.

**SUPPLEMENTAL DATA SHEET 1 |** Zipped text file containing the marker DNA sequences of length 25. The gene markers (column 2) are unique to their label (column 1), but the intergene and haplotype sequences are not.

**SUPPLEMENTAL DATA SHEET 2 |** Raw reads from synthetic evaluation, whose results are in **Table 2**. Each of the six simulated individuals have a pair of fastq files for each parental haplotype.

**SUPPLEMENTAL TABLE 1 |** Individual gene and haplotype-pair predictions. Columns A and B contain the family name and relationship. Column C contains KPI's haplotype-pair predictions, represented by a haplotype list. Each pair is separated with a '|'. For example, 'cA01~tA01+ cA01~tB01_2DS5|cA01~tB01_2DS5+cA01~tB05' means the prediction is either haplotype-pairs cA01~tA01 and cA01~tB01_2DS5 or haplotype-pairs cA01~tB01_2DS5 and cA01~tB05. The haplotype list in column C represents all possible haplotype-pairs fitting the presence-absence genotypes. Column D is a count of the number of haplotype-pair predictions in column C's haplotype list. Column E is the haplotype list in column C reduced by family relationships, and column F indicates whether or not it is different from the original haplotype list in column C. Column G is the original haplotype list in column C reduced by EM; its results should not be used as they have limited accuracy. Columns H-W are the gene presence-absence predictions.

## REFERENCES

1. Martin AM, Freitas EM, Witt CS, Christiansen FT. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* (2000) 51:268–80. doi: 10.1007/s002510050620
2. Martin AM, Kulski JK, Gaudieri S, Witt CS, Freitas EM, Trowsdale J, et al. Comparative genomic analysis, diversity and evolution of two KIR haplotypes A and B. *Gene* (2004) 335:121–31. doi: 10.1016/j.gene.2004.03.018
3. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol* (2013) 13:133–44. doi: 10.1038/nri3370
4. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Res* (2012) 22:1845–54. doi: 10.1101/gr.137976.112
5. Pyo C-W, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different Patterns of Evolution in the Centromeric and Telomeric Regions of Group A and B Haplotypes of the Human Killer Cell Ig-Like Receptor Locus. *PloS One* (2010) 5:e15115. doi: 10.1371/journal.pone.0015115
6. Manser AR, Weinhold S, Uhrberg M. Human KIR repertoires: shaped by genetic diversity and evolution. *Immunol Rev* (2015) 267:178–96. doi: 10.1111/imr.12316
7. Roe D, Vierra-Green C, Pyo C-W, Eng K, Hall R, Kuang R, et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun* (2017) 18:127–34. doi: 10.1038/gene.2017.10
8. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* (2014) 43(D1):D423–31. doi: 10.1093/nar/gku1161
9. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, et al. Human Diversity in Killer Cell Inhibitory Receptor Genes. *Immunity* (1997) 7:753–63. doi: 10.1016/S1074-7613(00)80394-5
10. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, et al. Imputation of KIR Types from SNP Variation Data. *Am J Hum Genet* (2015) 97:593–607. doi: 10.1016/j.ajhg.2015.09.005
11. Roe D, Vierra-Green C, Pyo C-W, Geraghty DE, Spellman SR, Maiers M, et al. A Detailed View of KIR Haplotype Structures and Gene Families as Provided by a New Motif-based Multiple Sequence Alignment. *Front Immunol* (2020). doi: 10.3389/fimmu.2020.585731
12. *dwgsim*. Available at: https://github.com/nh13/dwgsim.
13. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* (2014) 22:221–7. doi: 10.1038/ejhg.2013.118
14. Vierra-Green C, Roe D, Jayaraman J, Trowsdale J, Traherne J, Kuang R, et al. Estimating KIR Haplotype Frequencies on a Cohort of 10,000 Individuals: A Comprehensive Study on Population Variations, Typing Resolutions, and Reference Haplotypes. *PloS One* (2016) 11:e0163973. doi: 10.1371/journal.pone.0163973
15. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *Am J Hum Genet* (2016) 99:375–91. doi: 10.1016/j.ajhg.2016.06.023
16. *PING Github issue 5*. Available at: https://github.com/Hollenbach-lab/PING/issues/5.
17. dos Santos EJM, McCabe A, Gonzalez-Galarza FF, Jones AR, Middleton D. Allele Frequencies Net Database: Improvements for storage of individual genotypes and analysis of existing data. *Hum Immunol* (2016) 77:238–48. doi: 10.1016/j.humimm.2015.11.013
18. Chen J, Madireddi S, Nagarkar D, Migdal M, Heiden JV, Chang D, et al. *In silico* tools for accurate HLA and KIR inference from clinical

sequencing data empower immunogenetics on individual-patient and population scales. *Brief Bioinform* (2020) bbaa223. doi: 10.1093/bib/bbaa223

19. Wright PA. Killer-cell immunoglobulin-like receptor assessment algorithms in haemopoietic progenitor cell transplantation: current perspectives and future opportunities. *HLA* (2020) 95(5):435–48. doi: 10.1111/tan.13817
20. Roe D, Kuang R. Accurate and Efficient KIR Gene and Haplotype Inference from Genome Sequencing Reads with Novel K-mer Signatures. *bioRxiv* (2019). doi: 10.1101/541938
21. *Lifelines Biobank*. Available at: http://www.lifelines.nl.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Immunology

REVIEW
published: 14 December 2020
doi: 10.3389/fimmu.2020.601089

# From Chickens to Humans: The Importance of Peptide Repertoires for MHC Class I Alleles

Jim Kaufman[1,2]*

[1] School of Biological Sciences, Institute for Immunology and Infection Research, University of Edinburgh, Edinburgh, United Kingdom, [2] Department of Pathology, University of Cambridge, Cambridge, United Kingdom

OPEN ACCESS

**Edited by:**
Eric O. Long,
National Institutes of Health (NIH),
United States

**Reviewed by:**
Malcolm John Wyness Sim,
National Institutes of Health (NIH),
United States
Paul J. Norman,
University of Colorado Denver,
United States

***Correspondence:**
Jim Kaufman
jim.kaufman@ed.ac.uk

**Specialty section:**
This article was submitted to
NK and Innate Lymphoid
Cell Biology,
a section of the journal
Frontiers in Immunology

**Received:** 31 August 2020
**Accepted:** 30 October 2020
**Published:** 14 December 2020

**Citation:**
Kaufman J (2020) From Chickens to
Humans: The Importance of Peptide
Repertoires for MHC Class I Alleles.
Front. Immunol. 11:601089.
doi: 10.3389/fimmu.2020.601089

In humans, killer immunoglobulin-like receptors (KIRs), expressed on natural killer (NK) and thymus-derived (T) cells, and their ligands, primarily the classical class I molecules of the major histocompatibility complex (MHC) expressed on nearly all cells, are both polymorphic. The variation of this receptor-ligand interaction, based on which alleles have been inherited, is known to play crucial roles in resistance to infectious disease, autoimmunity, and reproduction in humans. However, not all the variation in response is inherited, since KIR binding can be affected by a portion of the peptide bound to the class I molecules, with the particular peptide presented affecting the NK response. The extent to which the large multigene family of chicken immunoglobulin-like receptors (ChIRs) is involved in functions similar to KIRs is suspected but not proven. However, much is understood about the two MHC-I molecules encoded in the chicken MHC. The BF2 molecule is expressed at a high level and is thought to be the predominant ligand of cytotoxic T lymphocytes (CTLs), while the BF1 molecule is expressed at a much lower level if at all and is thought to be primarily a ligand for NK cells. Recently, a hierarchy of BF2 alleles with a suite of correlated properties has been defined, from those expressed at a high level on the cell surface but with a narrow range of bound peptides to those expressed at a lower level on the cell surface but with a very wide repertoire of bound peptides. Interestingly, there is a similar hierarchy for human class I alleles, although the hierarchy is not as wide. It is a question whether KIRs and ChIRs recognize class I molecules with bound peptide in a similar way, and whether fastidious to promiscuous hierarchy of class I molecules affect both T and NK cell function. Such effects might be different from those predicted by the similarities of peptide-binding based on peptide motifs, as enshrined in the idea of supertypes. Since the size of peptide repertoire can be very different for alleles with similar peptide motifs from the same supertype, the relative importance of these two properties may be testable.

**Keywords: epistasis, Avian immunology, immunopeptidomics, B locus, BF-BL region, minimal essential major histocompatibility complex**

## INTRODUCTION

Molecules encoded by the major histocompatibility complex (MHC) of jawed vertebrates play central roles in immune responses as well as other important biological processes (1). Among these molecules are the classical class I molecules, which are defined by presentation of peptides on the cell surface, high and wide expression and high polymorphism. There are also non-classical class I molecules that lack one or more of these properties; in this report, only the classical class I molecules will be considered and will be abbreviated MHC-I.

MHC-I molecules bound to appropriate peptides on a cell surface are ligands for thymus-derived (T) lymphocytes through the T cell receptor (TCR) composed of α and β chains (along with the co-receptor CD8), with the outcome generally being death of the target cell through apoptosis (2). The cytotoxic T lymphocytes (CTLs) are important agents for response to infectious pathogens (particularly viruses) and cancers. The repertoire of TCRs is formed by somatic mutational mechanisms in individual cells and is vast and cross-reactive, so that in principle any MHC molecule bound to any peptide could be recognized (3). In fact, selection of T cells in the thymus strongly affects the TCR repertoire, but, to a first approximation, it is the polymorphism of the MHC molecules along with self-peptides that determines thymic selection, presentation of peptides, and thus immune responses (4).

However, many MHC-I molecules are also ligands for natural killer (NK) cells through a variety of NK receptors (NKRs), with the potential outcomes including cytokine release and target cytotoxicity (2). Analogous to T cell education based on the MHC molecules and self-peptides present in an individual, the responses of NK cells depend on the particular MHC molecules present during development, a phenomenon referred to as education, licencing, or tuning (5). Both NKRs and MHC-I ligands are polymorphic, with the interactions of particular receptors with particular ligands varying markedly in strength. Since the MHC and the regions encoding NKRs are located on different chromosomes, the genetic result is epistasis, which in humans and mice affects infectious disease, autoimmunity, and reproduction. Indeed, there appears to be antagonistic selection between immune responses and reproduction in humans (6).

MHC-I molecules (7) generally bind short peptides, 9–11 amino acids in length, along a groove between two α-helices above a β-pleated sheet. The peptides are tightly bound at the N- and C-termini by eight highly-conserved amino acids in pockets A and F, so that longer peptides bulge in the middle. Specificity of binding to different MHC-I alleles arises from peptide interactions with the polymorphic amino acids that line the groove, often with deeper pockets B and F being most important, but with other pockets being important in some alleles. The important pockets typically bind just one amino acid or a few amino acids with side chains that have very similar chemical properties, although some promiscuous pockets allow many different amino acids. The particular amino acids generally allowed to bind in the important pockets, the so-called anchor residues, give rise to peptide motifs for MHC-I alleles. Many alleles have been grouped into several supertypes (8) based on similarities in peptide motifs and in polymorphic amino acids lining the pockets. Some motifs are quite stringent in their requirements while others are more permissive, leading the concepts of fastidious and promiscuous MHC-I alleles with differently sized peptide repertoires (9).

TCRs recognize the side chains of peptide residues that point up and away from the peptide-binding groove, mostly in the middle of the peptide (10). It has long been known that the particular peptides bound to MHC-I molecules could influence interaction with inhibitory NKRs (11–14), which eventually was refined to NKR interaction with side chains near the end of the peptide (typically residues 7 and 8 of a 9mer) (15, 16). Moreover, both viral and bacterial peptides have been reported to affect recognition by activating NKRs (17, 18).

Among the questions that will be considered in this report are the extent to which the size of the peptide repertoire may influence the binding NKRs, and the extent to which MHC-I alleles within a supertype have the same sized peptide repertoire. In order to approach these questions, it is appropriate to review what is known about peptide repertoires, beginning with chicken class I molecules.

## THE CHICKEN MHC: A SIMPLE SYSTEM FOR DISCOVERY

The vast majority of what is known about the MHC and MHC molecules was discovered in humans and biomedical models like mice (1). In typical placental mammals (**Figure 1**), the MHC is several megabase pairs (Mbp) of DNA with hundreds of genes, separated into haplotype blocks by several centimorgans (cM) of recombination. The few MHC-I genes located in the class I region are separated from the few class II genes in the class II region by the class III region which contains many unrelated genes. Some genes involved in the class I antigen processing and presentation pathway (APP) are also located in the MHC, including two genes for inducible proteasome components (LMPs or PSMBs), two genes for the transporter for antigen presentation (TAP1 and TAP2) and the dedicated chaperone and peptide editor tapasin (TAPBP). However, these class I APP genes are located in the class II region and are more-or-less functionally monomorphic (19–21), working well for nearly all loci and alleles of MHC-I molecules. In humans, the three loci of MHC-I molecules may not be interchangeable: HLA-A and -B present peptides to CTLs with only some alleles acting as NKR ligands, while HLA-C is less well-expressed and mostly functions as an NKR ligand (22, 23). There is also evidence that HLA-A and -B may do different jobs, since HLA-B is more strongly associated with responses to rapidly evolving small (RNA) viruses, while HLA-A may be more involved with large double-stranded DNA viruses (24).

In contrast, the chicken MHC is small and simple (**Figure 1**), and evolves mostly as stable haplotypes (9, 25). The BF-BL region of the B locus is less than 100 kB and contains two MHC-I genes (BF1 and BF2) flanking the TAP1 and TAP2 genes, with the TAPBP gene sandwiched between two class II B genes

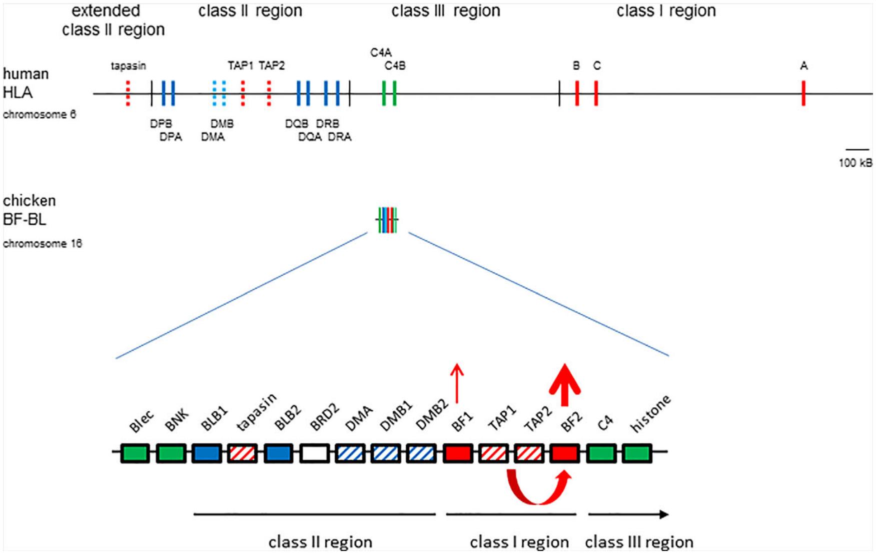


**FIGURE 1** | The chicken MHC (BF-BL region) is smaller and simpler than the human MHC (HLA locus), with a single dominantly-expressed MHC-I molecule due to co-evolution with peptide-loading genes. Colored vertical lines or boxes indicate genes, with names above; thin vertical lines indicate region boundaries, with names above or below; location is roughly to scale, with the length of approximately 100 kB indicated. Thickness of arrows pointing up indicate level of expression, co-evolution between the TAP genes and the BF2 class I gene indicated by a curved arrow beneath the genes. Genes from the class I system, red; the class II system, blue; the class III or other regions, green; solid colors indicate classical genes while striped colors indicate genes involved in peptide loading. Figure from (9).

nearby, and with the class III region on the outside. There is evidence only for historic recombination within this region, with no examples of recombinants from over 20,000 informative progeny in deliberate mating, although there is clear recombination just outside (in the so-called TRIM and BG regions) (26–28). As a result, alleles of these strongly-linked genes stay together for long periods of time, so that the APP genes are all highly polymorphic and co-evolve with the BF2 gene (9, 29). As an example, the peptide translocation specificity of the TAP is appropriate for the peptide binding specificity of the BF2 molecule encoded by that haplotype (30, 31). Apparently as a result, the BF2 molecule is far more expressed and also more polymorphic than the BF1 molecule (32, 33). Thus far, the evidence is that the BF2 molecule presents peptides to CTLs, while BF1 functions as a ligand for NK cells (34).

This simplicity of the chicken MHC can make it easier to discover phenomena that are difficult to discern in the more complicated MHC of humans and other placental mammals. For example, there are many examples of strong genetic associations of the B locus (and in some cases, the BF-BL region) with responses to economically-important diseases, including Marek's disease caused by an oncogenic herpesvirus, infectious bronchitis caused by a coronavirus and avian influenza (9, 35). In contrast, the strongest associations of the human MHC are with autoimmune diseases, with the strongest associations with infectious disease being with small viruses like HIV (1). One hypothesis for this perceived difference is the fact that the human MHC has a multigene family of class I molecules which confer more-or-less resistance to most viral pathogens (reading out as weak genetic associations), while the chicken MHC has a single dominantly-expressed class I molecule, which either finds a protective peptide or not (reading out as strong genetic associations) (9, 36).

Other examples of discovery from the apparent simplicity of the chicken MHC will be described below, but it has become clear that other aspects of the avian immunity may be very complex, for instance the chicken NKR system.

## PROMISCUOUS AND FASTIDIOUS CLASS I ALLELES IN CHICKENS

One of the discoveries that was facilitated by the presence of a single dominantly-expressed chicken class I molecule is an apparent inverse correlation between peptide repertoire and cell surface expression, along with strong correlations with resistance to infectious diseases. Some so-called promiscuous BF2 alleles bind a wide variety of peptides but have a relatively low expression on the cell surface cell, while other so-called fastidious BF2 alleles bind a much more limited variety of peptides but have higher cell surface expression (9, 32, 37, 38).

It is not clear whether there is a hierarchy or two general groups of alleles, or to what extent the cell surface expression levels are exactly an inverse of the peptide repertoire. The analysis of expression level by flow cytometry is quantitative, but the exact levels vary for different cell types. The peptide repertoires are far more difficult to quantify, with even immunopeptidomics that fairly accurately counts numbers of different peptides by mass spectrometry suffering from the drawback that the abundance of any given peptide is laborious to establish definitively. However, for certain well-studied standard B haplotypes, the peptide-motifs based on gas phase sequencing and on immunopeptidomics, as well as the pockets defined by crystal structures, give qualitative rationales for the peptide repertoires (9, 32, 37–41). The peptide translocation specificities of the TAP alleles from the few haplotypes examined provide additional support (30, 31).

The high expressing fastidious alleles typically bind peptides through three positions with only one or a few amino acids allowed (32, 39–41). For instance, the BF2 allele from the B4 haplotype (BF2*004:01) binds almost entirely octamer peptides with three acidic residues: Asp or Glu at positions P2 and P5, and Glu (with very low levels of hydrophobic amino acids) at position P8, which fits the basic amino acids forming the so-called pockets B, C, and F in wire models and the crystal structure. BF2*012:01 binds octamer peptides with Val or Ile at position P5 and Val at position P8, but with a variety of amino acids at position P2, which is an anchor residue as seen by structure. BF2*015:01 binds peptides with Arg or Lys in position P1, Arg in position P2 and Tyr (with very low levels of Phe and Trp) at positions P8 or P9. In fact, these BF2 alleles with fastidious motifs can bind a wider variety of peptides *in vitro* than are actually found on the cell surface (31, 39); the TAP translocation specificities are more restrictive than the BF2 peptide binding specificities.

In contrast, it would appear that a variety of binding mechanisms can lead to low expressing alleles with promiscuous motifs. BF2*021:01 has certain positions with small amino acids leading to a wide bowl in the centre of the binding groove, within which Asp24 and Arg9 can move, remodelling the binding site to accommodate a wide variety of 10mer and 11mer peptides with co-variation of P2 and Pc-2 (two from the end), along with hydrophobic amino acids at the final position. Interactions between P2, Pc-2, Asp24, and Arg9 allow a wide range of amino acid side chains in the peptide, with at least three major modes of binding (37, 38). Analysis of peptide translocation in B21 cells shows the specificity is less stringent than the BF2*021:01 molecule (31). In another mechanism, BF2*002:01 binds peptides with two hydrophobic pockets for P2 and Pc, but the pockets are wide and shallow, allowing a variety of small to medium-sized amino acid side chains (38). BF2*014:01 also has two pockets, accommodating medium to large-sized amino acid side chains at P2 and positive charge(s) at Pc (38). Binding many different hydrophobic amino acids allows a promiscuous motif, since hydrophobic amino acids are so common in proteins.

Another interesting feature of chicken class I molecules is C-terminal overhang of peptides outside of the groove. In placental mammals, one of the eight invariant residues that bind the peptide N- and C-termini is Tyr84, which blocks the egress of the peptide at the C-terminus. However, in chickens (and all other jawed vertebrates outside of placental mammals), the equivalent residue is an Arg (42, 43) and this change allows the peptide to hang out of the groove, as has been found in crystal structures of BF2*012:01 and 014:01 (28, 30). At least one low expressing class I allele with an otherwise fastidious motif shows lots of such overhangs (C. Tregaskes, R. Martin and J. Kaufman, unpublished), suggesting that the TAP translocation specificity (or perhaps the TAPBP peptide editing) controls the extent to which overhangs are permitted. Interestingly, the equivalent position in class II molecules is also Arg, allowing most peptides to hang out of the groove, with some of these overhangs recognized by TCRs (40, 43, 44). Thus, the presence of such overhangs may be a third mechanism for chicken class I promiscuity, and may affect both TCR and NKR recognition, as do peptide sidechains within the groove in humans (10, 16).

The reason for the inverse correlation of peptide repertoire with cell surface expression is not clear. Among the possibilities are biochemical mechanisms, which are highlighted by the fact that all chicken BF2 alleles have nearly identical promoters, and that the amount of protein inside the cell does not differ much, but that the amount that moves to the cell surface is more for fastidious than promiscuous alleles (31). Thus, the amount of time associated with the TAPBP and TAP in the peptide-loading complex (PLC) could be a mechanistic reason. Another potential biochemical mechanism might be stability and degradation; promiscuous alleles from cells are overall less stable than fastidious alleles in solution, but pulse-chase experiments of *ex vivo* lymphoctyes show no obvious difference in turn-over (31). As a second reason, the correlation could arise from the need to balance effective immune responses to pathogens and tumours with the potential for immunopathology and autoimmunity. A third possibility is the need to balance negative selection in the thymus with the production of an effective naïve TCR repertoire: more peptides presented would mean more T cells would be deleted, but since TCR signal depends on the number of peptide-MHC complexes, lower class I expression would mean fewer T cells would be deleted (9, 45). If true, the expression level would be the important property, since it would mirror the need for an effective T cell repertoire.

What makes this inverse correlation so interesting is the association with resistance and susceptibility to economically-important pathogens. A correlation with low cell surface expression was first noticed for resistance to the tumours arising from the oncogenic herpesvirus that causes Marek's disease, and later understood to correlate with a wide peptide repertoire (9, 36–38). Important caveats include the fact that the association of the B locus with resistance to Marek's disease, while still true for experimental lines, have not been found for current commercial chickens (46–48); an explanation may be the fact that poultry breeders have strongly enriched for low expressing class I alleles in their flocks so that the MHC no longer has a differential effect (C. Tregaskes, R. Martin and J. Kaufman, unpublished). Another caveat may be that there are various measures of the progress of Marek's disease, and the BF-BL region correlations may not be the same for all of them. A third caveat is that the BF-BL region is composed of strongly-linked genes, so that the gene (or genes) responsible for resistance are not yet definitively identified; an example is the evidence for the effect of the BG1 gene (49). An important counter to these caveats is that there is evidence that MHC haplotypes with low-expressing class I alleles confer resistance to other infectious viral diseases, including Rous sarcoma, infectious bronchitis and avian influenza (9, 50–52). Importantly, there is little recognized evidence that the high expressing alleles provide important immune benefit to chickens.

## PROMISCUOUS AND FASTIDIOUS MHC MOLECULES IN HUMANS: GENERALISTS AND SPECIALISTS

Having clear evidence of the inverse correlation of cell surface expression level with peptide repertoire of chicken BF2 alleles and infectious disease resistance, it was natural to ask whether these relationships are fundamental properties of class I molecules, as opposed to some special feature of chicken class I molecules. Such evidence for human HLA-A and B alleles with hints towards potential mechanisms was not hard to find.

Progression of human immunodeficiency virus (HIV) infection to frank acquired immunodeficiency disease syndrome (AIDS) is one of the best examples for an association of infectious disease with the human MHC. Some HLA alleles lead to fast progression and death, while others result in very slow progression, for which the individuals can be called elite controllers (53, 54). The number of peptides from the human proteome predicted to bind four HLA-B alleles was compared to odds ratio for AIDS, finding that the most fastidious alleles were the most protective. Although the correlation with disease resistance was the reverse of what was found for chickens (45), flow cytometric analyses of these four alleles on *ex vivo* blood lymphoid and myeloid cells showed that these human class I molecules had the same inverse correlation between peptide repertoire and cell surface expression as in chickens (38).

A mechanism of resistance by such elite controlling HLA-B alleles has been reported: the presentation of particular HIV peptides to CTLs which the virus can mutate to escape the immune response, but only at the cost of much reduced viral fitness. For such alleles, the virus is caught between a rock and a hard place (55, 56). The protection to the human host afforded by binding and presenting such special peptides led to a hypothesis (9, 38), in which the promiscuous class I alleles act as generalists, providing protection against many common and slowly evolving pathogens (as in chickens), while the fastidious alleles act as specialists, with particular alleles providing protection against a given new and quickly evolving pathogen (as in humans). There are some caveats to this story. One is that the predictions are only a reflection of reality, based on benchmarking the predictions made by such algorithms against experimental data from immunopeptidomics (57). Another is that other explanations are possible; a study calculating the number of peptides predicted for class II alleles concluded that promiscuous alleles would appear based on the number of pathogens in particular environments (58).

Another study determined the number of peptides from dengue virus predicted to bind 27 common HLA-A and -B alleles, concluding that there is a wide variation in peptide repertoire that is inversely correlated with stability (59), similar to what was found for chicken class I molecules. Three of the four HLA-B alleles analyzed in the human proteome study were also analyzed in this dengue study and followed the same hierarchy (**Figure 2**). Interestingly, more HLA-B alleles were found at the fastidious end of the spectrum and more HLA-A alleles were found at the promiscuous end, particularly HLA-A2 variants. It would appear that HLA-A and B alleles have a range of peptide repertoires but perhaps not as wide as in chickens. The fastidious chicken class I molecules typically have three fastidious anchor residues compared to two for human class I molecules, while the promiscuous HLA-A2 variants each allow two or three hydrophobic amino acids compared to five or more for BF2*002:01. Unlike chicken MHC-I molecules, peptide overhangs from human MHC-I molecules are relatively rare and require major re-adjustments of peptide-binding site, such as movement of α-helices that line the groove (60, 61), so this is not likely to be a general mechanism for promiscuity in humans.

## MECHANISMS FOR ESTABLISHING PEPTIDE REPERTOIRES IN HUMAN CLASS I MOLECULES

The question arises whether the peptide motif determines the peptide repertoire for human class I molecules, given that the APP genes for human class I molecules are more-or-less functionally monomorphic so that all class I alleles will get a

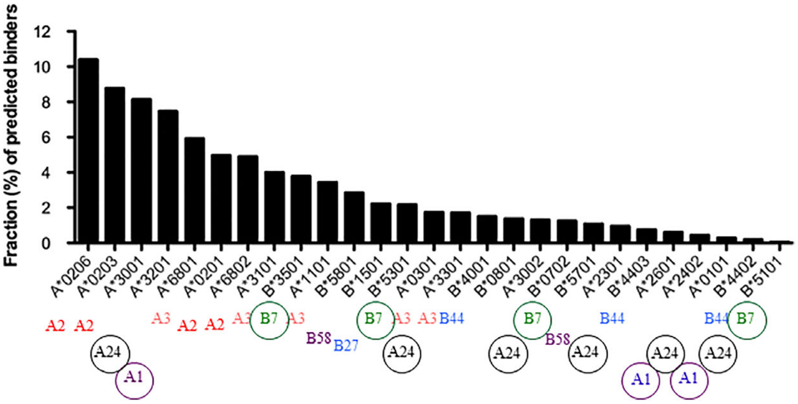


**FIGURE 2** | The predictive peptide repertoires for 27 common HLA-A and -B alleles [from (59), Copyright 2013. The American Association of Immunologists, Inc.] compared to the supertypes of these alleles [from (8)] show that the peptide motifs do not correlate well with peptide repertoire for some supertypes.

wide and promiscuous set of peptides and peptide editing. As mentioned above, supertypes of MHC-I molecules have been defined based on shared peptide motifs and on amino acids lining the pockets of peptide binding sites (8). A comparison of the peptide repertoires presented in the Dengue study (59) with such supertypes (**Figure 2**) shows that some peptide motifs correlate well with peptide repertoire (for example A2, A3, etc); for example, the alleles falling within the A2 supertype are all found at the promiscuous end of the repertoire. However, alleles from other supertypes (for example A1, A24, and B7), are found across the spectrum of repertoires. Thus peptide motifs do not equate with peptide repertoires, giving the possibility of discriminating between the two designations in terms of contribution towards disease.

A study on the dependence of cell surface expression of HLA-B alleles on TAPBP (also known as tapasin) may give a clue as to the discrepancy between peptide motif and peptide repertoire (62). There are many reports of particular pairs of alleles varying in TAPBP-dependence, and positions in the α2 and α3 domains have been identified that affect this dependence. A hierarchy of dependence was described for 27 HLA-B alleles (63), and a rough correlation with the hierarchy of peptide repertoire was found: fastidious alleles were by-and-large more dependent of TAPBP for cell surface expression, while promiscuous alleles were not (9). Such dependence would fit with the stability of class I molecules mentioned above: Peptide editing by TAPBP leads to the fastidious class I molecules retaining only the peptides that have the highest affinity, while promiscuous class I molecules would bind and move to the cell surface with any peptide with a minimal affinity. Moreover, the authors concluded that tapasin-independent alleles were linked to more rapid progression from HIV infection to death from AIDS (63).

Interestingly, this dependence of TAPBP correlated with the ease of refolding with peptides *in vitro* (in the absence of TAPBP), with both human and chicken promiscuous alleles refolding more easily (38, 62). Whether chicken class I alleles have the same dependence *in vivo* is not yet clear, since TAPBP is highly polymorphic, with the TAPBP and BF2 alleles in each haplotype likely to have co-evolved (64).

## HLA-C AND BF1: FLIES IN THE OINTMENT?

The fact that there are relationships of cell surface expression, peptide repertoire and resistance to infection disease both for BF2 in chickens and for HLA-A and -B in humans suggested that these are fundamental properties of MHC-I molecules. However, the evidence for HLA-C in humans and BF1 in chickens, which have some intriguing similarities, may not fit this emerging paradigm (9).

HLA-C is the result of an ancient gene duplication of HLA-B, but the two differ in several important ways (22, 23). Both HLA-B and -C molecules are polymorphic, are up-regulated upon inflammation, and bind and present peptides to αβ T cells. However, HLA-B molecules are expressed at the RNA, protein and cell surface levels as well as HLA-A molecules. HLA-B molecules are major CTL ligands on virally-infected cells, but some alleles carrying the Bw4 epitope on the α1 helix of the peptide-binding domain are also recognized by NKRs, specifically the killer immunoglobulin receptors with three extracellular domains (3D KIRs).

In contrast, HLA-C molecules are expressed at a low RNA level and are found at about 10% of the level of HLA-A or -B molecules on the surfaces of cells where all three loci are expressed. However, they are also expressed on extravillous trophoblasts (EVT) in the absence of HLA-A and -B molecules. HLA-C alleles are known as important NKR ligands, by carrying either C1 or C2 epitopes on the α1 helix of the peptide-binding domain, which are recognized by different KIRs with two extracellular domains (2D KIRs). Moreover, different HLA-C alleles have different RNA and cell surface protein levels, for which those with higher expression are correlated with slow progression from HIV infection to AIDs, and with some evidence to suggest that this correlation is due to recognition by CTLs (65, 66). There have been no experiments reported to explicitly test the relationship of peptide repertoire and cell surface expression of HLA-C alleles, but the determination of cell surface expression has been reported to be very complex, including effects of promoters, miRNA, assembly, stability and peptide-binding specificity (67).

Much less is known about the chicken BF1 gene, but it has some similarities to the HLA-C. BF1 molecules are expressed at a much lower level than BF2 molecules, at the level of RNA, protein and antigenic peptide (32, 33). There are far fewer alleles of BF1 than BF2, with ten-fold less BF1 RNA found in most haplotypes and with some haplotypes missing a BF1 gene altogether peptide (32, 33). BF1 is also thought to be primarily an NKR ligand (34), and most BF1 alleles carry a C1 motif on the α1 helix of the peptide-binding domain (68, 69). Examination of sequences suggests that most BF1 alleles have similar peptide-binding grooves, with the few examples of other sequences likely to have been due to sequence contributions from the BF2 locus (C. Tregaskes, R. Martin and J. Kaufman, unpublished). An unsolved question is how BF1 alleles interact effectively with the highly polymorphic TAP and TAPBP alleles, for instance accommodating the very different peptides from translocated by TAPs in different haplotypes. Perhaps the typical BF1 molecule is highly promiscuous, but there are few data for either peptide repertoire or cell surface expression among BF1 alleles.

## THE OTHER SIDE OF THE COIN: RECEPTORS ON NATURAL KILLER CELLS

An enormous body of scientific literature describes the very complex evolution, structure and function of NKRs and NK cells in primates and mice (2, 70, 71). Two kinds of NKRs are found in humans, lectin-like receptors found in the natural killer complex (NKC) and the KIRs in the leukocyte receptor complex (LRC). The KIRs are a highly polymorphic multigene family with copy number variation, and share the human LRC with other immunoglobulin-like receptors, including leukocyte immunoglobulin-like receptors (LILRs) and a single receptor for antibodies (Fcμ/αR or CD351).

Some of these transmembrane receptors have cytoplasmic tails with immune-tyrosine inhibitory motifs (ITIMs), others have basic residues in the transmembrane region which allow association with signaling chains bearing immune-tyrosine activating motifs (ITAMs), and a few have both. The polymorphic NKRs interact with polymorphic MHC-I molecules, 2D KIRs with HLA-C and 3D KIRs with certain HLA-A and HLA-B alleles. As mentioned above, the interactions of the particular alleles present in LRC and MHC, which are on different chromosomes, lead to differing outcomes, which read out as genetic epistasis with effects on immunity, autoimmunity and reproduction.

In chickens, almost all of the known immunoglobulin-like receptors related to KIRs are found on a single microchromosome, different from the one on which is found the MHC (72, 73). These chicken immunoglobulin-like receptors (ChIRs) include those with activating, inhibitory and both motifs (ChIR-A, -B, and -AB), and 1D, 2D, and 4D extracellular regions. Sequencing studies suggest there can be haplotypes with few ChIR genes in common, suggesting both copy number variation and polymorphism (74–76). However, a gene typing method for 1D domains suggested relatively stable haplotypes, with only some examples of recombination during matings (77). The only molecules that have clear functions are many ChIR-AB molecules that bind IgY, the antibody isotype that acts somewhat like IgG in mammals (78–80). It seems very likely that there are both activating and inhibitory NKRs among these ChIRs, but thus far no data for NKR function. Whether such putative NKRs recognize BF1, BF2, or both is as yet unknown, and whether there is epistasis between the ChIR and MHC microchromosomes is untested.

Among the lectin-like NKR genes located in the NKC in humans and mice are one or more NKR-P1 genes (also known as NK1.1, KRLB1, or CD161) paired with the lectin-like ligands (LLT1 in humans and Clr in mice). In chickens, there are only two lectin-like genes located in the region syntenic to the NKC, and neither of those appears to encode NKRs; one is expressed mainly in thrombocytes (81, 82). However, there is a pair of NKR-P1/ligand genes in the chicken MHC (25, 83), known as BNK (sometimes identified as Blec1) and Blec (sometimes identified as Blec2). The receptor encoded by the highly polymorphic BNK gene was assumed to interact with the nearly monomorphic Blec gene, but a reporter cell line with one BNK allele was found not to respond to BF1, BF2 or Blec, but to spleen cells bearing a trypsin sensitive ligand (84, 85). A trypsin-sensitive ligand on a particular chicken cell line was found to reproduce the result with the reporter cells, but the nature of that ligand remains unknown (E. K. Meziane, B. Viertlboeck, T. Göbel and J. Kaufman, unpublished). Possibilities include other lectin-like genes in the BG region or the Y region of the MHC microchromosome (28, 86).

The effect of peptide repertoire of class I molecules on NK recognition has not carefully examined in either humans or chickens, but some speculations may be worth considering. A wider peptide repertoire may increase the number (although unlikely the proportion) of peptides with appropriate amino acids to affect binding to KIRs and ChIRs, both at the level of response and potentially at the level of education (licensing or tuning), including the recently described phenomenon of cis-tuning (87). However, the increase in breadth of peptide repertoire may be balanced by the decrease of cell surface expression of the class I molecules, which may mean that peptide repertoire may not exert an enormous effect on inhibitory NK responses. In contrast, any increase in peptide repertoire may allow additional pathogen peptides to be recognized by activating NKRs. A special consideration are C-terminal overhangs, which may be particularly frequent in at least some alleles of chicken class I molecules. Such C-terminal overhangs in human class II molecules can directly affect T cell recognition (44), so it is possible that NKR interactions could also be affected.

## CONCLUSIONS

The simplicity of the chicken MHC has allowed discoveries of phenomena that were harder to discern from analysis of the more complicated MHC of humans and mice (such as the existence of promiscuous and fastidious MHC-I alleles), but comparison between the immune systems of chickens and mammals has been fruitful (as in the development of the generalist-specialist hypothesis). For human MHC-I molecules, peptide motifs (as identified by supertypes) can be separated from peptide repertoire (as defined thus far by peptide prediction), but their impact on NKR recognition has not been tested. Moreover, careful analysis of Pc-1 and Pc-2 residues in promiscuous versus fastidious alleles with respect to peptide repertoire has not yet been carried for either humans or chickens. Given that the most basic understanding of NKR recognition in chickens has yet to gained, the importance of C-terminal peptide overhang from chicken MHC-I alleles for NKR recognition or NK function has not yet been assessed. Thus, it is clear that there is much work to do to understand NK cell function in chickens, and how that function relates to what is known in typical mammals including humans and mice.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## FUNDING

Wellcome Trust Investigator Award for Science 110106/A/15/Z, currently awarded to University of Edinburgh, which supports my labs in Edinburgh and Cambridge.

## ACKNOWLEDGMENTS

I thank the Wellcome Trust for support (Investigator Award 110106/Z/15/Z), Drs. Clive Tregaskes, Samer Halabi and Kat Gordon for critical reading of the manuscript, and particularly two reviewers for very helpful suggestions.

## REFERENCES

1. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet* (2013) 14:301–23. doi: 10.1146/annurev-genom-091212-153455
2. Djaoud Z, Parham P. HLAs, TCRs, and KIRs, a triumvirate of human cell-mediated immunity. *Annu Rev Biochem* (2020) 89:717–39. doi: 10.1146/annurev-biochem-011520-102754
3. Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol* (2012) 12 (9):669–77. doi: 10.1038/nri3279
4. Starr TK, Jameson SC, Hogquist KA. Positive and negative selection of T cells. *Annu Rev Immunol* (2003) 21:139–76. doi: 10.1146/annurev.immunol.21.120601.141107
5. Long EO, Kim HS, Liu D, Peterson ME, Rajagopalan S. Controlling natural killer cell responses: integration of signals for activation and inhibition. *Annu Rev Immunol* (2013) 31:227–58. doi: 10.1146/annurev-immunol-020711-075005
6. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol* (2013) 13(2):133–44. doi: 10.1038/nri3370
7. Yaneva R, Schneeweiss C, Zacharias M, Springer S. Peptide binding to MHC class I and II proteins: new avenues from new methods. *Mol Immunol* (2010) 47(4):649–57. doi: 10.1016/j.molimm.2009.10.008
8. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. *BMC Immunol* (2008) 9:1. doi: 10.1186/1471-2172-9-1
9. Kaufman J. Generalists and Specialists: A new view of how MHC class I molecules fight infectious pathogens. *Trends Immunol* (2018) 39(5):367–79. doi: 10.1016/j.it.2018.01.001
10. Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. T cell antigen receptor recognition of antigen-presenting molecules. *Annu Rev Immunol* (2015) 33:169–200. doi: 10.1146/annurev-immunol-032414-112334
11. Malnati MS, Peruzzi M, Parker KC, Biddison WE, Ciccone E, Moretta A, et al. Peptide specificity in the recognition of MHC class I by natural killer cell clones. *Science* (1995) 267(5200):1016–8. doi: 10.1126/science.7863326
12. Hilton HG, McMurtrey CP, Han AS, Djaoud Z, Guethlein LA, Blokhuis JH, et al. The intergenic recombinant HLA-B*46:01 has a distinctive peptidome that includes KIR2DL3 ligands. *Cell Rep* (2017) 19(7):1394–405. doi: 10.1016/j.celrep.2017.04.059
13. Peruzzi M, Wagtmann N. Long EO. A p70 killer cell inhibitory receptor specific for several HLA-B allotypes discriminates among peptides bound to HLA-B*2705. *J Exp Med* (1996) 184(4):1585–90. doi: 10.1084/jem.184.4.1585
14. Sim MJ, Malaker SA, Khan A, Stowell JM, Shabanowitz J, Peterson ME, et al. Canonical and cross-reactive binding of NK cell inhibitory receptors to HLA-C allotypes Is dictated by peptides bound to HLA-C. *Front Immunol* (2017) 8:193. doi: 10.3389/fimmu.2017.00193
15. Saunders PM, Vivian JP, O'Connor GM, Sullivan LC, Pymm P, Rossjohn J. Brooks AG. A bird's eye view of NK cell receptor interactions with their MHC class I ligands. *Immunol Rev* (2015) 267(1):148–66. doi: 10.1111/imr.12319
16. Das J, Khakoo SI. NK cells: tuned by peptide? *Immunol Rev* (2015) 267 (1):214–27. doi: 10.1111/imr.12315
17. Naiyer MM, Cassidy SA, Magri A, Cowton V, Chen K, Mansour S, et al. KIR2DS2 recognizes conserved peptides derived from viral helicases in the context of HLA-C. *Sci Immunol* (2017) 2(15):eaal5296. doi: 10.1126/sciimmunol.aal5296
18. Sim MJW, Rajagopalan S, Altmann DM, Boyton RJ, Sun PD, Long EO. Human NK cell receptor KIR2DS4 detects a conserved bacterial epitope presented by HLA-C. *Proc Natl Acad Sci USA* (2019) 116(26):12964–73. doi: 10.1073/pnas.1903781116
19. Momburg F, Roelse J, Howard JC, Butcher GW, Hämmerling GJ, Neefjes JJ. Selectivity of MHC-encoded peptide transporters from human, mouse and rat. *Nature* (1994) 367(6464):648–51. doi: 10.1038/367648a0
20. Obst R, Armandola EA, Nijenhuis M, Momburg F, Hämmerling GJ. TAP polymorphism does not influence transport of peptide variants in mice and humans. *Eur J Immunol* (1995) 25(8):2170–6. doi: 10.1002/eji.1830250808
21. Williams AP, Bevan S, Bunce M, Houlston R, Welsh KI, Elliott T. Identification of novel Tapasin polymorphisms and linkage disequilibrium to MHC class I alleles. *Immunogenetics* (2000) 52(1-2):9–11. doi: 10.1007/s002510000244
22. Blais ME, Dong T, Rowland-Jones S. HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? *Immunology* (2011) 133(1):1–7. doi: 10.1111/j.1365-2567.2011.03422.x
23. Anderson SK. Molecular evolution of elements controlling HLA-C expression: Adaptation to a role as a killer-cell immunoglobulin-like receptor ligand regulating natural killer cell function. *HLA* (2018) 92(5):271–8. doi: 10.1111/tan.13396
24. Hertz T, Nolan D, James I, John M, Gaudieri S, Phillips E, et al. Mapping the landscape of host-pathogen coevolution: HLA class I binding and its relationship with evolutionary conservation in human and viral proteins. *J Virol* (2011) 85(3):1310–21. doi: 10.1128/JVI.01966-10
25. Kaufman J, Milne S, Göbel TW, Walker BA, Jacob JP, Auffray C, et al. The chicken B locus is a minimal essential major histocompatibility complex. *Nature* (1999) 401(6756):923–5. doi: 10.1038/44856
26. Hála K, Chaussé AM, Bourlet Y, Lassila O, Hasler V, Auffray C. Attempt to detect recombination between B-F and B-L genes within the chicken B complex by serological typing, in vitro MLR, and RFLP analyses. *Immunogenetics* (1988) 28(6):433–8. doi: 10.1007/BF00355375
27. Fulton JE, McCarron AM, Lund AR, Pinegar KN, Wolc A, Chazara O, et al. Miller MM. A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex B region between BG2 and CD1A1. *Genet Sel Evol* (2016) 48:1. doi: 10.1186/s12711-015-0181-x
28. Salomonsen J, Chattaway JA, Chan AC, Parker A, Huguet S, Marston DA, et al. Sequence of a complete chicken BG haplotype shows dynamic expansion and contraction of two gene lineages with particular expression patterns. *PLoS Genet* (2014) 10(6):e1004417. doi: 10.1371/journal.pgen.1004417
29. Kaufman J, Jacob J, Shaw I, Walker B, Milne S, Beck S, et al. Gene organisation determines evolution of function in the chicken MHC. *Immunol Rev* (1999) 167:101–17. doi: 10.1111/j.1600-065x.1999.tb01385.x
30. Walker BA, Hunt LG, Sowa AK, Skjødt K, Göbel TW, Lehner PJ, et al. The dominantly expressed class I molecule of the chicken MHC is explained by coevolution with the polymorphic peptide transporter (TAP) genes. *Proc Natl Acad Sci USA* (2011) 108(20):8396–401. doi: 10.1073/pnas.1019496108
31. Tregaskes CA, Harrison M, Sowa AK, van Hateren A, Hunt LG, Vainio O, et al. Surface expression, peptide repertoire, and thermostability of chicken class I molecules correlate with peptide transporter specificity. *Proc Natl Acad Sci USA* (2016) 113(3):692–7. doi: 10.1073/pnas.1511859113
32. Wallny HJ, Avila D, Hunt LG, Powell TJ, Riegert P, Salomonsen J, et al. Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens. *Proc Natl Acad Sci USA* (2006) 103(5):1434–9. doi: 10.1073/pnas.0507386103
33. Shaw I, Powell TJ, Marston DA, Baker K, van Hateren A, Riegert P, et al. Different evolutionary histories of the two classical class I genes BF1 and BF2 illustrate drift and selection within the stable MHC haplotypes of chickens. *J Immunol* (2007) 178(9):5744–52. doi: 10.4049/jimmunol.178.9.5744
34. Kim T, Hunt HD, Parcells MS, van Santen V, Ewald SJ. Two class I genes of the chicken MHC have different functions: BF1 is recognized by NK cells while BF2 is recognized by CTLs. *Immunogenetics* (2018) 70(9):599–611. doi: 10.1007/s00251-018-1066-2
35. Miller MM, Taylor RLJr. Brief review of the chicken Major Histocompatibility Complex: the genes, their distribution on chromosome 16, and their contributions to disease resistance. *Poult Sci* (2016) 95(2):375–92. doi: 10.3382/ps/pev379
36. Kaufman J, Völk H, Wallny HJ. A "minimal essential Mhc" and an "unrecognized Mhc": two extremes in selection for polymorphism. *Immunol Rev* (1995) 143:63–88. doi: 10.1111/j.1600-065x.1995.tb00670.x
37. Koch M, Camp S, Collen T, Avila D, Salomonsen J, Wallny HJ, et al. Structures of an MHC class I molecule from B21 chickens illustrate promiscuous peptide binding. *Immunity* (2007) 27(6):885–99. doi: 10.1016/j.immuni.2007.11.007
38. Chappell P, Meziane el K, Harrison M, Magiera Ł, Hermann C, Mears L, et al. Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife* (2015) 4:e05345. doi: 10.7554/eLife.05345

39. Zhang J, Chen Y, Qi J, Gao F, Liu Y, Liu J, et al. Narrow groove and restricted anchors of MHC class I molecule BF2*0401 plus peptide transporter restriction can explain disease susceptibility of B4 chickens. *J Immunol* (2012) 189(9):4478–87. doi: 10.4049/jimmunol.1200885
40. Xiao J, Xiang W, Zhang Y, Peng W, Zhao M, Niu L, et al. An Invariant Arginine in Common with MHC Class II Allows Extension at the C-Terminal End of Peptides Bound to Chicken MHC Class I. *J Immunol* (2018) 201 (10):3084–95. doi: 10.4049/jimmunol.1800611
41. Li X, Zhang L, Liu Y, Ma L, Zhang N, Xia C. Structures of the MHC-I molecule BF2*1501 disclose the preferred presentation of an H5N1 virus-derived epitope. *J Biol Chem* (2020) 295(16):5292–306. doi: 10.1074/jbc.RA120.012713
42. Kaufman J, Andersen R, Avila D, Engberg J, Lambris J, Salomonsen J, et al. Different features of the MHC class I heterodimer have evolved at different rates. Chicken B-F and beta 2-microglobulin sequences reveal invariant surface residues. *J Immunol* (1992) 148(5):1532–46.
43. Kaufman J, Salomonsen J, Flajnik M. Evolutionary conservation of MHC class I and class II molecules–different yet the same. *Semin Immunol* (1994) 6 (6):411–24. doi: 10.1006/smim.1994.1050
44. Zavala-Ruiz Z, Strug I, Walker BD, Norris PJ, Stern LJ. A hairpin turn in a class II MHC-bound peptide orients residues outside the binding groove for T cell recognition. *Proc Natl Acad Sci USA* (2004) 101(36):13279–84. doi: 10.1073/pnas.0403371101
45. Kosmrlj A, Read EL, Qi Y, Allen TM, Altfeld M, Deeks SG, et al. Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* (2010) 465(7296):350–4. doi: 10.1038/nature08997
46. Heifetz EM, Fulton JE, O'Sullivan NP, Arthur JA, Wang J, Dekkers JC, et al. Mapping quantitative trait loci affecting susceptibility to Marek's disease virus in a backcross population of layer chickens. *Genetics* (2007) 177(4):2417–31. doi: 10.1534/genetics.107.080002
47. Heifetz EM, Fulton JE, O'Sullivan NP, Arthur JA, Cheng H, Wang J, et al. Mapping QTL affecting resistance to Marek's disease in an F6 advanced intercross population of commercial layer chickens. *BMC Genomics* (2009) 10:20. doi: 10.1186/1471-2164-10-20
48. Wolc A, Arango J, Jankowski T, Settar P, Fulton JE, O'Sullivan NP, et al. Genome-wide association study for Marek's disease mortality in layer chickens. *Avian Dis* (2013) 57(2 Suppl):395–400. doi: 10.1637/10409-100312-Reg.1
49. Goto RM, Wang Y, Taylor RLJr, Wakenell PS, Hosomichi K, Shiina T, et al. BG1 has a major role in MHC-linked resistance to malignant lymphoma in the chicken. *Proc Natl Acad Sci USA* (2009) 106(39):16740–5. doi: 10.1073/pnas.0906776106
50. Khare VM, Saxena VK, Tomar A, Singh KP, Singh KB, Tiwari AK. MHC-B haplotypes impact susceptibility and resistance to RSV-A infection. *Front Biosci* (2018) 10:506–19. doi: 10.2741/e837
51. Boonyanuwat K, Thummabutra S, Sookmanee N, Vatchavalkhu V, Siripholvat V. Influences of major histocompatibility complex class I haplotypes on avian influenza virus disease traits in Thai indigenous chickens. *Anim Sci J* (2006) 77:285–9. doi: 10.1111/j.1740-0929.2006.00350.x
52. Banat GR, Tkalcic S, Dzielawa JA, Jackwood MW, Saggese MD, Yates L, et al. Association of the chicken MHC B haplotypes with resistance to avian coronavirus. *Dev Comp Immunol* (2013) 39(4):430–7. doi: 10.1016/j.dci.2012.10.006
53. Goulder PJ, Walker BD. HIV and HLA class I: an evolving relationship. *Immunity* (2012) 37(3):426–40. doi: 10.1016/j.immuni.2012.09.005
54. Carrington M, Walker BD. Immunogenetics of spontaneous control of HIV. *Annu Rev Med* (2012) 63:131–45. doi: 10.1146/annurev-med-062909-130018
55. Schneidewind A, Brockman MA, Yang R, Adam RI, Li B, Le Gall S, et al. Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. *J Virol* (2007) 81(22):12382–93. doi: 10.1128/JVI.01543-07
56. Miura T, Brockman MA, Schneidewind A, Lobritz M, Pereyra F, Rathod A, et al. HLA-B57/B*5801 human immunodeficiency virus type 1 elite controllers select for rare gag variants associated with reduced viral replication capacity and strong cytotoxic T-lymphocyte recognition. *J Virol* (2009) 83(6):2743–55. doi: 10.1128/JVI.02265-08
57. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, Nielsen M, et al. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. *PLoS Comput Biol* (2020) 16(5): e1007757. doi: 10.1371/journal.pcbi.1007757
58. Manczinger M, Boross G, Keményě L, Möller V, Lenz TL, Papp B, et al. Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLoS Biol* (2019) 17(1):e3000131. doi: 10.1371/journal.pbio.3000131
59. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J Immunol* (2013) Dec 15191(12):5831–9. doi: 10.4049/jimmunol.1302101
60. McMurtrey C, Trolle T, Sansom T, Remesh SG, Kaever T, Bardet W, et al. Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. *Elife* (2016) 5:e12556. doi: 10.7554/eLife.12556
61. Remesh SG, Andreatta M, Ying G, Kaever T, Nielsen M, McMurtrey C, et al. Unconventional peptide presentation by Major Histocompatibility Complex (MHC) class I allele HLA-A*02:01: BREAKING CONFINEMENT. *J Biol Chem* (2017) 292(13):5262–70. doi: 10.1074/jbc.M117.776542
62. Rizvi SM, Salam N, Geng J, Qi Y, Bream JH, Duggal P, et al. Distinct assembly profiles of HLA-B molecules. *J Immunol* (2014) 192(11):4967–76. doi: 10.4049/jimmunol.1301670
63. Raghavan M, Geng J. HLA-B polymorphisms and intracellular assembly modes. *Mol Immunol* (2015) 68(2 Pt A):89–93. doi: 10.1016/j.molimm.2015.07.007
64. van Hateren A, Carter R, Bailey A, Kontouli N, Williams AP, Kaufman J, et al. A mechanistic basis for the co-evolution of chicken tapasin and major histocompatibility complex class I (MHC I) proteins. *J Biol Chem* (2013) 288(45):32797–808. doi: 10.1074/jbc.M113.474031
65. Blais ME, Zhang Y, Rostron T, Griffin H, Taylor S, Xu K, et al. High frequency of HIV mutations associated with HLA-C suggests enhanced HLA-C-restricted CTL selective pressure associated with an AIDS-protective polymorphism. *J Immunol* (2012) 188(9):4663–70. doi: 10.4049/jimmunol.1103472
66. Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, et al. Influence of HLA-C expression level on HIV control. *Science* (2013) 340(6128):87–91. doi: 10.1126/science.1232685
67. Kaur G, Gras S, Mobbs JI, Vivian JP, Cortes A, Barber T, et al. Structural and regulatory diversity shape HLA-C protein expression levels. *Nat Commun* (2017) 8:15924. doi: 10.1038/ncomms15924
68. Ewald SJ, Livant EJ. Distinctive polymorphism of chicken B-FI (major histocompatibility complex class I) molecules. *Poult Sci* (2004) 83(4):600–5. doi: 10.1093/ps/83.4.600
69. Livant EJ, Brigati JR, Ewald SJ. Diversity and locus specificity of chicken MHC B class I sequences. *Anim Genet* (2004) 35(1):18–27. doi: 10.1111/j.1365-2052.2003.01078.x
70. Lanier LL. NK cell recognition. *Annu Rev Immunol* (2005) 23:225–74. doi: 10.1146/annurev.immunol.23.021704.115526
71. Barrow AD, Trowsdale J. The extended human leukocyte receptor complex: diverse ways of modulating immune responses. *Immunol Rev* (2008) 224:98–123. doi: 10.1111/j.1600-065X.2008.00653.x
72. Viertlboeck BC, Göbel TW. The chicken leukocyte receptor cluster. *Vet Immunol Immunopathol* (2011) 144(1-2):1–10. doi: 10.1016/j.vetimm.2011.07.001
73. Straub C, Neulen ML, Sperling B, Windau K, Zechmann M, Jansen CA, et al. Chicken NK cell receptors. *Dev Comp Immunol* (2013) 41(3):324–33. doi: 10.1016/j.dci.2013.03.013
74. Laun K, Coggill P, Palmer S, Sims S, Ning Z, Ragoussis J, et al. The leukocyte receptor complex in chicken is characterized by massive expansion and diversification of immunoglobulin-like Loci. *PLoS Genet* (2006) 2(5):e73. doi: 10.1371/journal.pgen.0020073
75. Lochner KM, Viertlboeck BC, Göbel TW. The red junglefowl leukocyte receptor complex contains a large, highly diverse number of chicken immunoglobulin-like receptor (CHIR) genes. *Mol Immunol* (2010) 47(11-12):1956–62. doi: 10.1016/j.molimm.2010.05.001
76. Viertlboeck BC, Gick CM, Schmitt R, Du Pasquier L, Göbel TW. Complexity of expressed CHIR genes. *Dev Comp Immunol* (2010) 34(8):866–73. doi: 10.1016/j.dci.2010.03.007
77. Meziane EK, Potts ND, Viertlboeck BC, Løvlie H, Krupa AP, Burke TA, et al. Bi-functional chicken immunoglobulin-like receptors with a single extracellular domain (ChIR-AB1): potential framework genes among a

relatively stable number of genes per haplotype. *Front Immunol* (2019) 10:2222. doi: 10.3389/fimmu.2019.02222

78. Viertlboeck BC, Schweinsberg S, Hanczaruk MA, Schmitt R, Du Pasquier L, Herberg FW, et al. The chicken leukocyte receptor complex encodes a primordial, activating, high-affinity IgY Fc receptor. *Proc Natl Acad Sci USA* (2007) 104(28):11718–23. doi: 10.1073/pnas.0702011104
79. Arnon TI, Kaiser JT, West APJr, Olson R, Diskin R, Viertlboeck BC, et al. The crystal structure of CHIR-AB1: a primordial avian classical Fc receptor. *J Mol Biol* (2008) 381(4):1012–24. doi: 10.1016/j.jmb.2008.06.082
80. Viertlboeck BC, Schweinsberg S, Schmitt R, Herberg FW, Göbel TW. The chicken leukocyte receptor complex encodes a family of different affinity FcY receptors. *J Immunol* (2009) 182(11):6985–92. doi: 10.4049/jimmunol.0803060
81. Chiang HI, Zhou H, Raudsepp T, Jesudhasan PR, Zhu JJ. Chicken CD69 and CD94/NKG2-like genes in a chromosomal region syntenic to mammalian natural killer gene complex. *Immunogenetics* (2007) 59(7):603–11. doi: 10.1007/s00251-007-0220-z
82. Neulen ML, Göbel TW. Identification of a chicken CLEC-2 homologue, an activating C-type lectin expressed by thrombocytes. *Immunogenetics* (2012) 64(5):389–97. doi: 10.1007/s00251-011-0591-z
83. Rogers SL, Göbel TW, Viertlboeck BC, Milne S, Beck S, Kaufman J. Characterization of the chicken C-type lectin-like receptors B-NK and B-lec suggests that the NK complex and the MHC share a common ancestral region. *J Immunol* (2005) 174(6):3475–83. doi: 10.4049/jimmunol.174.6.3475
84. Rogers SL, Kaufman J. High allelic polymorphism, moderate sequence diversity and diversifying selection for B-NK but not B-lec, the pair of lectin-like receptor genes in the chicken MHC. *Immunogenetics* (2008) 60 (8):461–75. doi: 10.1007/s00251-008-0307-1
85. Viertlboeck BC, Wortmann A, Schmitt R, Plachý J, Göbel TW. Chicken C-type lectin-like receptor B-NK, expressed on NK and T cell subsets, binds to a ligand on activated splenocytes. *Mol Immunol* (2008) 45(5):1398–404. doi: 10.1016/j.molimm.2007.08.024
86. Rogers S, Shaw I, Ross N, Nair V, Rothwell L, Kaufman J, et al. Analysis of part of the chicken Rfp-Y region reveals two novel lectin genes, the first complete genomic sequence of a class I alpha-chain gene, a truncated class II beta-chain gene, and a large CR1 repeat. *Immunogenetics* (2003) 55(2):100–8. doi: 10.1007/s00251-003-0553-1
87. Goodson-Gregg FJ, Krepel SA, Anderson SK. Tuning of human NK cells by endogenous HLA-C expression. *Immunogenetics* (2020) 72(4):205–15. doi: 10.1007/s00251-020-01161-x

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Immunology

REVIEW
published: 08 March 2021
doi: 10.3389/fimmu.2021.601518

# Spondyloarthritis and the Human Leukocyte Antigen (HLA)-B*27 Connection

*Chengappa G. Kavadichanda [1*†], Jie Geng [2†], Sree Nethra Bulusu [1], Vir Singh Negi [1] and Malini Raghavan [2*]*

[1] *Department of Clinical Immunology, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India,* [2] *Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, United States*

Heritability of Spondyloarthritis (SpA) is highlighted by several familial studies and a high association with the presence of human leukocyte antigen (HLA)-B*27. Though it has been over four decades since the association of HLA-B*27 with SpA was first determined, the pathophysiological roles played by specific HLA-B*27 allotypes are not fully understood. Popular hypotheses include the presentation of arthritogenic peptides, triggering of endoplasmic reticulum (ER) stress by misfolded HLA-B*27, and the interaction between free heavy chains or heavy chain homodimers of HLA-B*27 and immune receptors to drive IL-17 responses. Several non-HLA susceptibility loci have also been identified for SpA, including endoplasmic reticulum aminopeptidases (ERAP) and those related to the IL-23/IL-17 axes. In this review, we summarize clinical aspects of SpA including known characteristics of gut inflammation, enthesitis and new bone formation and the existing models for understanding the association of HLA-B*27 with disease pathogenesis. We also examine newer insights into the biology of HLA class I (HLA-I) proteins and their implications for expanding our understanding of HLA-B*27 contributions to SpA pathogenesis.

**Keywords: HLA-B*27, spondyloarthritis, ER stress, free heavy chain, IL-23/IL-17 axis, ERAP1**

## INTRODUCTION

Spondyloarthritis (SpA) is a group of seronegative arthritides, which includes ankylosing spondylitis (AS), psoriatic arthritis (PsA), reactive arthritis (ReA), undifferentiated SpA and enteropathy related arthritis (EA) (1). The global prevalence of SpA ranges from 0.2 to 1.61% in the general population. The numbers depend on the geographic area, the study population, data sources and the case definition used to classify SpA, which has evolved considerably over the years (2). The above subtypes of SpA share several phenotypic characteristics (**Figure 1**), which include inflammatory lesions in the axial and peripheral joints, enthesitis (inflammation at the insertion sites of tendons and ligaments into the bone), uveitis (inflammation in the eye) and enteritis (inflammation in the small intestine), in varying combinations and frequencies. Clinically, individuals with SpA present with low back ache, alternating gluteal pain and stiffness of the spine, all of which worsen with rest and improve upon exercise. Along with these axial symptoms, individuals with SpA also have inflamed peripheral joints and entheseal sites. The skeletal manifestations are often associated with several extra-articular features in the eye, skin and gut, depending on the subtype of SpA. The features start off insidiously and progress chronically

OPEN ACCESS

***Edited by:***
*Brian Duncan Tait,*
*Australian Red Cross Blood Service, Australia*

***Reviewed by:***
*Ozgur Kasapcopur,*
*Istanbul University-Cerrahpasa, Turkey*
*Uma Kanga,*
*All India Institute of Medical Sciences, India*

***Correspondence:**
*Chengappa G. Kavadichanda*
*doc.chengappa@gmail.com*
*Malini Raghavan*
*malinir@umich.edu*

[†] *These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to Alloimmunity and Transplantation, a section of the journal Frontiers in Immunology*

***Received:*** *01 September 2020*
***Accepted:*** *08 February 2021*
***Published:*** *08 March 2021*

***Citation:***
*Kavadichanda CG, Geng J, Bulusu SN, Negi VS and Raghavan M (2021) Spondyloarthritis and the Human Leukocyte Antigen (HLA)-B*27 Connection. Front. Immunol. 12:601518. doi: 10.3389/fimmu.2021.601518*

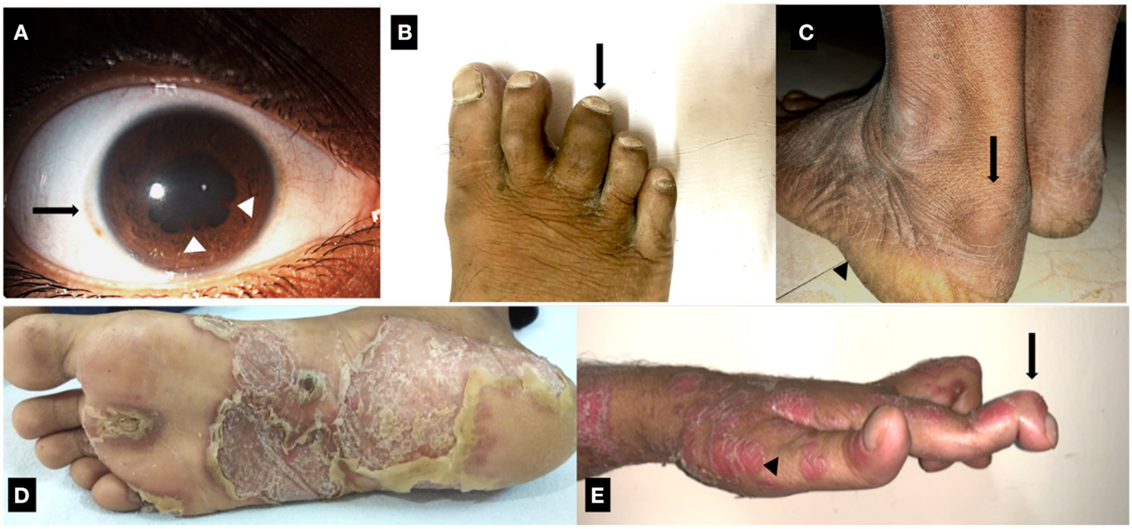


**FIGURE 1 |** Clinical manifestations of Spondyloarthritis. **(A)** Resolving recurrent acute anterior uveitis of right eye in a case of ankylosing spondylitis. Solid arrow: mild circum-corneal congestion. Arrowhead: posterior synechiae. **(B)** Dactylitis of third toe of the right foot in a case of Psoriatic arthritis. **(C)** Enthesitis involving the Achilles tendon in a case of ankylosing spondylitis. Solid arrow: retrocalcaneal bursitis, which co-occurs with Achilles enthesitis. Arrowhead: site of plantar fasciitis.
**(D)** Keratoderma blennorrhagica involving the sole of a patient with Reactive arthritis. **(E)** Psoriasis (arrowhead) with deforming peripheral arthritis of hand joints. Solid arrow: arthritis and deformity of distal interphalangeal joint.

in most of the subtypes of SpA except in ReA. ReA presents with an acute onset of ankle or knee inflammation along with dactylitis (combined inflammation of the joint and soft tissues in fingers or toes), enthesitis, conjunctivitis and skin lesions like keratoderma blennorhagicum (mucous-laden skin lesions) (**Figure 1**) and circinate balanitis (skin inflammation around the glans penis) (3). PsA is an SpA phenotype which occurs in individuals with psoriatic skin lesions. It is characterized by a variable combination of sacroiliitis (inflammation of the sacroiliac joints), which is usually asymmetrical, oligo (affecting 2–4 joints) to polyarthritis (affecting > 4 joints), enthesitis, dactylitis (**Figure 1**) and chronic anterior or posterior uveitis (4). AS is considered as the prototype SpA and it presents with bilateral symmetric sacroiliitis, acute anterior uveitis, peripheral arthritis and enthesitis (5). SpA can have varying degrees of bowel inflammation ranging from microscopic asymptomatic colitis to overt inflammatory bowel disease (IBD) in 7–10% of AS and PsA (6). Besides this, 3% of patients with IBD have AS, 10% have subclinical sacroiliac joint involvement and 30% of the cases have SpA-like musculoskeletal symptoms (7).

The skeletal manifestations progress from inflammatory lesions to a combination of erosive, destructive and proliferative pathology. The proliferative pathology, characterized by new bone formation, progresses partly independent of the inflammatory process. In particular, the axial skeletal involvement progresses to cause severe disability as a result of irreversible fusion of vertebral bodies and formation of marginal syndesmophytes (calcification and new bone formation in ligaments). This results in a stiff spine referred to as the "bamboo spine" even in patients whose inflammation is fairly controlled. The reason for this progression is still not elucidated and remains as an important unanswered question in the management of SpA.

In inflammatory arthritic diseases, such as Rheumatoid Arthritis (RA), the commonly used and most effective immunosuppressive agents are glucocorticoids (GC) and conventional synthetic disease-modifying anti-rheumatic drugs (csDMARDS). For unknown reasons, these drugs have negligible benefit in SpA. On the other hand, non-steroidal anti-inflammatory drugs (NSAIDs) have excellent efficacy in relieving pain and spinal stiffness in SpA. Besides NSAIDs, monoclonal antibodies against tumor necrosis factor alpha (TNFα) and IL-17 are the current standard of care in SpA (8). Even though these biological agents reduce inflammation, it remains unclear if

**Abbreviations:** ALK 2, Activin receptor-like kinase-2; ANKENT, Ankylosing enthesopathy; AS, Ankylosing Spondylitis; AxSpA, Axial Spondyloarthrits; BMP, Bone morphogenic protein; csDMARDs, Conventional synthetic diseases modifying anti-rheumatic drugs; DBA, Dilute, Brown and non-Agouti; ER, Endoplasmic reticulum; ERA, Enthesitis related arthritis; ERAP, Endoplasmic Reticulum Aminopeptidase; FDR, First degree relatives; FHC, Free heavy chain; GC, Glucocorticoids; HLA, Human leukocyte antigen; hβ2m, Human beta2-microglobulin; IBD, Inflammatory Bowel Diseases; IL, Interleukin; ILC, Innate-like lymphocyte; JAK, Janus Kinase; jPsA, Juvenile Psoriatic arthritis; jSpA, Juvenile Spondyloarthritis; KIR, Killer cell immunoglobulin-like receptor; LILR, Leukocyte immunoglobulin (Ig)-like receptors; MAIT, Mucosal-associated invariant T; MHC, Major histocompatibility complex; MSC, Mesenchymal stem cells; NPEPPS, Aminopeptidase Puromycin Sensitive; NSAIDs, Non-steroidal anti-inflammatory drugs; PBMCs, Peripheral blood mononuclear cells; PIR, Paired immunoglobulin-like receptors; PLC, Peptide-loading complex; PsA, Psoriatic arthritis; PSpA, Peripheral Spondyloarthritis; RA, Rheumatoid arthritis; RARB, Retinoic acid receptor-β; ReA, Reactive arthritis; SNP, Single-nucleotide polymorphism; SpA, Spondyloarthritis; TAP, Transporter associated with antigen processing; TCR, T cell receptor; TGF, Transforming growth factor; TNAP, Tissue-nonspecific alkaline phosphatase; TNF, Tumor necrosis factor; UPR, Unfolded protein response; XBP1, X-box–binding protein 1.

**TABLE 1 |** Association of HLA-B*27 and various classes of Spondyloarthritis.

| Diseases | HLA-B*27 frequency | Comments |
|---|---|---|
| Ankylosing Spondylitis (AS) | 75–90% (12) | Prototypical SpA, which is classified based on the modified New York criteria, heavily banks on X-ray changes of sacroiliac joints. |
| Non-Radiographic Spondyloarthritis (Nr-SpA) | 75–90% (11,13) | This subclass was earlier classified as undifferentiated SpA. |
| Reactive arthritis (ReA) | 30–60% (3) | Few reports have linked HLA-B*27 with chronicity of ReA. |
| Psoriatic Arthritis (PsA) | 20–50% (18) | HLA-B*27 positive patients with psoriatic arthritis have higher incidence of enthesitis, dactylitis and symmetric sacroiliitis. |
| Enteropathy/Inflammatory bowel diseases related arthritis (EA/ IBD-SpA) | 10–40% (7) | HLA-B*27 is likely to increase the likelihood of having axSpA features in those with inflammatory bowel disease. |
| Enthesitis related arthritis (ERA) | 50–80% (21) | The juvenile counterpart of spondyloarthritis. They have predominant oligo arthritis in the beginning and a few patients gradually progress to have axial symptoms in the early adulthood. |
| Undifferentiated peripheral Spondyloarthritis (USpA) | 25–70% | This class is now limited to enteropathy related arthritis without overt features of IBD or PsA before the onset of psoriatic skin lesions. |

these drugs can reduce new bone formation. The response to the cytokine-targeted therapies are not uniform and there are a substantial number of patients who do not respond to either of these drugs. It is unclear who will respond to each drug and there is no convincing basis for one drug choice over the other. It is likely that numerous factors including genetic factors have a role in determining the differences in the onset, progress and response to treatment in SpA.

As discussed below, the association between SpA and HLA-B*27 is one of the strongest known associations between an HLA allele and disease. Several studies have highlighted the importance of this association for diagnosis, predicting disease phenotype and prognosis of SpA. However, our current understanding of the pathophysiologic pathways linking HLA-B*27 and SpA is still incomplete. In this review, we summarize the clinical associations between HLA-B*27 in patients with SpA and among their family members. We also examine the role of HLA-B*27 in enthesitis, new bone formation and gut pathology encountered in SpA. Finally, we elaborate on the unique properties of HLA-B*27 and highlight the newer findings related to HLA-B*27 biology which may partially explain the relevance of HLA-B*27 to SpA.

## HLA-B*27 AND DISEASE PHENOTYPE OF SpA

The occurrence of AS in a first degree relative (FDR) or other family members of the proband was known much before the identification of HLA-B*27-SpA association (9). The strength of association between HLA-B*27 and various SpA categories are variable (**Table 1**). The presence of HLA-B*27 results in the onset of AS symptoms at an early age (10, 11). The presence of HLA-B*27 also determines the distribution of inflammation across various organs in SpA. Uveitis, hip arthritis and sacroiliitis are more common among HLA-B*27$^+$ individuals (12), whereas the presence of HLA-B*27 in AS is negatively associated with peripheral arthritis and dactylitis (11).

Recent reports studying the broader classification of SpA, which include the non-radiographic phenotype, have found higher disease activity and worse functional scores in the B*27$^-$ patients (11, 13) suggesting that HLA-B*27$^-$ patients are likely to have worse disease at baseline, possibly as a result of delay in diagnosis. The presence of HLA-B*27 is likely to result in a prolonged course of illness as reflected by radiographic damage. Studies involving MRI imaging of sacroiliac joints have shown higher structural damage and inflammatory edema among HLA-B*27$^+$ patients (14). Moreover, when followed over several years, the progress in structural damage of sacroiliac joints and the propensity to develop marginal symmetric syndesmophytes along the vertebral column are higher among the HLA-B*27$^+$ individuals (15, 16).

HLA-B*27 is associated with onset of arthritis in individuals who developed psoriasis after 40 years of age (17). The presence of HLA-B*27 is also associated with bilateral sacroiliitis which otherwise is asymmetrical and unilateral in PsA (18). Inflammation in the sacroiliac joints detected by MRI is higher in HLA-B*27$^+$ PsA patients, which is similar to that seen in AS (18). In EA, presence of HLA-B*27 is associated with higher incidence of lower limb arthritis (19).

SpA in children and adolescents are currently classified either as enthesitis related arthritis (ERA) and juvenile psoriatic arthritis (jPsA) or as juvenile Spondyloarthritis (jSpA). The presence of HLA-B*27 among these individuals predisposes to development of skeletal deformities (20). Unlike their adult counterparts, patients with ERA have lower limb-predominant oligoarthritis, and late onset of axial involvement (21).

Overall, the data associate HLA-B*27 with uveitis and early onset disease, with radiologically severe axial manifestation in AS. The non-AS SpA group is predisposed to develop axial and peripheral in the presence of HLA-B*27. Among the jSpA, HLA-B*27 appears to be associated with a poorer prognosis.

## Gut Inflammation in SpA

SpA is characterized by inflammation in mainly four tissues, either in isolation or in combination. The gut, entheses, anterior chamber of the eye, and axial skeleton including the sacroiliac, intervertebral, costotransverse and facet joints are the tissues predominantly affected. The gut and entheses are proposed as the sites where inflammation is initiated (22, 23). Evidence garnered from histo/immuno-pathological studies, animal studies, and *in vitro* cell-based studies involving these tissues have furthered our understanding of SpA and are summarized here.

The epithelial barrier along with the mucosa-associated lymphoid tissue and the microbiota of an individual controls the delicate equilibrium between immune tolerance and activation. The gut involvement in SpA can range from overt IBD, to subclinical microscopic colitis, to dysbiosis in almost all patients with SpA (6, 24). Early experiments demonstrated that HLA-B*27 transgenic mice failed to develop SpA phenotype when reared in a germ-free environment (25). Reports have also shown that HLA-B*27 may alter the composition of the gut microbiota. *Bacteroides vulgatus*, for instance, is abundantly present in the HLA-B*27 transgenic Lewis lines as compared to the wild-type controls (26). Similarly, studies on gut biopsies of humans with AS have demonstrated higher colonies of certain bacterial communities (*Lachnospiraceae*, *Rikenellaceae, Porphyromonadaceae*, and *Bacteroidaceae*) (27) and a strong positive correlation of genus *Dialister* (28) and *Ruminococcus gnavus* (27, 29) with disease activity. The latter study also showed significant differences in microbiota composition between HLA-B*27$^+$ and HLA-B*27$^-$ siblings of AS (29), further suggesting a role for B*27 in determining the composition of microbiome in humans. These findings indicate that HLA-B*27 may have a role in altering immune responses by modifying the gut microbiome. The development of gut inflammation in HLA-B*27 transgenic rats can be altered by administering oral antibiotics. This treatment also reduces IL-1α and CCL2 levels and the number of Lin$^-$CD172a$^+$CD43$^{low}$ monocytes, subsets shown to have osteoclastogenic potential (30). A recent study investigating the role of metabolites in the gut of HLA-B*27 transgenic rat found that HLA-B*27 expression alters the intestinal metabolome of rats even before the onset of SpA symptoms. Moreover, administration of microbial metabolite propionate could attenuate development of SpA phenotype in these rats (31). It is, however, still not clear if the dysbiosis is due to inflammation or vice-versa.

Dysbiosis and presence of invasive bacteria in gut of AS alters the gut epithelial and gut vascular barrier along with dysregulated zonulin and tight junction expression. The leakiness due to altered tight junction in the gut results in increased levels of zonulin and bacterial products such as lipopolysaccharide (LPS), LPS-binding protein (BP), and intestinal fatty acid-BP in the serum of patients with AS (32) which can influence the differentiation of circulating monocytes.

The IL-23 and IL-17 pathways are linked to many autoimmune and auto-inflammatory diseases including AS (33–35). Genetic studies have suggested links between IL-23R, autoimmunity (33), gut inflammation (36) and AS (34), and some studies have reported increased IL-23 expression in AS (37, 38). Overexpression of IL-23 in the gut is thought to be a marker of intestinal inflammation, with Paneth cells (PC) being a major source of IL-23 (37). Studies with HLA-B*27 transgenic rats have demonstrated links between HLA-B*27 misfolding, the unfolded protein response (UPR) and IL-23 hyper-production (39). Other studies have indicated that misfolding of HLA-B*27 induces autophagy in the gut and downstream regulation of the production of IL-23 in AS (40). IL-23 can then activate Th17 cells, ILC3 cells, mucosal-associated invariant T (MAIT) cells, and γδ T cells involved in type 3 immunity (41). Overall, it seems that HLA-B*27, dysbiosis and activation of the IL-23/IL-17 axis at the gut are primal for inducing inflammation at remote sites including the joints and enthesis. However, the presence of all the SpA features including gut inflammation in HLA-B*27$^-$ SpA suggests that the precise role of B*27 in regulating the IL-23/IL-17 axis needs to be further explored. Other possible links between HLA-B*27 and IL-23/IL-17 axis will be further discussed below.

## Enthesitis in SpA

The enthesis is a fibrocartilaginous part of a ligament or tendon that inserts to the bone surface. Besides acting as a structure anchoring the muscle, tendon or capsule to the bone, the enthesis is a complex organ that efficiently transmits mechanical forces from muscles to bones across joints. Chronic inflammation of the entheseal complex (enthesitis) is a common clinical occurrence in SpA. In patients with SpA, enthesitis in the lower limbs (Achilles tendon and plantar fascia) is more common than that in upper limbs, probably due to the microtrauma burden at the former sites. In the TNF$^{\Delta ARE}$ mouse, which lack on-off regulation of TNF biosynthesis (42), enthesitis and new bone formation were promoted only upon inducing biomechanical stress (43), suggesting a key role for mechanical stress in the pathogenesis of SpA. Indeed, many sites subject to high mechanical stress, including the aortic root, the ciliary body of the eye, skin extensor surfaces, and the lung apex are target sites in the SpA group of diseases (44).

IL-23 seems to play an important role in precipitating enthesitis and resulting bone remodeling in SpA. Sherlock et al. found that entheseal sites of mice expressing IL-23 contained a population of CD3$^+$CD4$^-$CD8$^-$IL-23R$^+$, RAR-related orphan receptor γt (ROR-γt)$^+$ T cells. They further demonstrated that IL-23 overexpression was sufficient to precipitate enthesitis which predated the structural changes in joints, induced by prolonged inflammation. The IL-23 mediated pathology was dependent on the presence of CD3$^+$CD4$^-$CD8$^-$IL-23R$^+$ROR-γt$^+$ T cells and was independent of the presence of CD4$^+$ T cells including the Th17 cells. IL-23 induced expression of TNFα, IL-17 and IL-22 by the newly described subset of entheseal T cells (45). Furthermore, IL-22 was the dominant effector cytokine driving bone remodeling at the site of entheseal inflammation (45).

The clinical trial results of anti-IL-23 and anti-IL-17 in SpA indicate different effects. While anti-IL-17 agents had good outcomes in patients with axSpA (46, 47), anti-IL-23 agents failed to show benefits over placebo (48, 49). On the other hand, IL-23 inhibitors were beneficial in treating enthesitis in PsA (50), but the efficacy in improving arthritis and axial symptoms was not as impressive as those with anti-TNFs (46). These results have led to several untested hypotheses about the role of IL-23 in SpA. IL-23 may be an initiator for IL-17A and TNFα release, and axSpA may represent a chronic and mature form with a predominant IL-17-related phenotype. The other hypothesis is that the pathways driving the axial inflammation and peripheral inflammation have distinct etiologies and that IL-23 predominantly drives the peripheral SpA phenotype.

Another important finding in SpA is the detection of innate-like lymphocyte 3 (ILC 3) cells and γδ T cells at entheseal sites. Human entheses are shown to contain ILC3, based on expression of RORγt and IL-23R, and these cells induced IL-17A transcripts upon stimulation with IL-23 and IL-1β (51). These experiments were carried out on entheseal tissues of healthy individuals who had spinal surgeries and whose HLA-B*27 status was not known. It remains to be elucidated whether and how HLA-B*27 and chronic inflammation will alter this response. γδ T cells are also known to be a major source of IL-17 and TNFα. Experiments using mouse models and human enthesis have demonstrated that subsets of entheseal γδ T-cells may be maintained as self-renewing tissue resident cells. These cells upregulate IL-17A production in an IL-23-independent (52) or dependent (53) manner when activated, making them important players in local inflammation. Even though an earlier study in humans showed higher levels of IL-23R$^+$ γδ T cells in the peripheral blood of HLA-B*27$^+$ AS patients (54), their role as entheseal resident cells in humans is yet to be established. Biomechanical or unknown infective stressors are likely the triggers for enthesitis. These stressors possibly trigger IL-23-mediated responses at the entheseal site, which is self-perpetuating in some cases leading to widespread inflammation. More precise information about the entheseal resident cell populations, their interactions with HLA-B*27 and their role in injury and healing may provide us with better insights into the pathophysiology of enthesitis in SpA.

## New Bone Proliferation

In SpA, the process of intense inflammation and repair is accompanied by formation of bony spicules from the underlying trabecular bone. New bone formation is not limited to HLA-B*27-mediated disease. Non-HLA-B*27 transgenic animal models like the Dilute, Brown and non-Agouti (DBA) 1 (55), Ankylosing enthesopathy (ANKENT) mice (56), and mouse models with overexpression of IL-23 show evidence of new bone formation, while TNF overexpressing models like the TNF$^{\Delta ARE}$ mice do not show osteoproliferation (43, 45). Data from long term follow-up of patients treated with anti-TNF agents have not been conclusive about prevention of osteoproliferation (57). These findings suggest that there are factors beyond inflammation that are likely to result in new bone formation. The presence of HLA-B*27 appears to increase the severity and magnitude of osteoproliferation in AS (15, 16) and PsA (18). Inflammation, neoangiogenesis, and new bone formation are visualized in clinical practice even in asymptomatic entheseal sites of SpA patients using ultrasound power doppler, MRI, and radiographs, and are HLA-B*27-dependent (58). The process of new bone formation in AS appears to be a result of trans-differentiation followed by ossification of cartilage and direct bone formation (59). The common understanding is that new bone formation is facilitated by bone morphogenic proteins (BMPs) and Wnt proteins. In DBA/1 mice, new bone formation was driven by BMP signaling, which was inhibited by Noggin, a BMP antagonist (60). Studies in humans using immunohistochemistry of synovial tissue showed higher expression of BMP-2 and BMP-6 in inflamed SpA tissue, but not non-inflamed tissue (61). These genes were up-regulated in fibroblast-like synoviocytes by IL-1β and TNFα, which are important cytokines in SpA pathogenesis. Another study in AS patients showed abnormal secretion of Noggin and BMP2, by mesenchymal stem cells (MSC), causing an imbalance that is suggested to induce abnormal osteogenic differentiation (62). The IL-17 and IL-22 accumulated at the enthesis during acute inflammation may also facilitate the proliferation, migration and osteogenic differentiation of MSCs (63).

As noted above, Sherlock et al. attributed new bone formation to be IL-23 and IL-22 driven (45). Some recent studies have shed light on a potential direct pathway linking HLA-B*27 and osteoproliferation. Expression of HLA-B*27:04 and B*27:05 but not B*07:02 along with human beta2-microglobulin (hβ2m) in *Drosophila* resulted in the loss of cross veins in the wings. This phenotype resulted from a dominant-negative effect on the BMP receptor saxophone, and elevated phosphorylation of a *Drosophila* receptor mediated Smad. The human saxophone ortholog ALK2 and HLA-B*27 were shown to interact in lymphoblastoid cells from AS patients. Active ALK2 inhibits BMP signaling via phosphorylation of Smads. HLA-B*27-mediated inhibition of ALK2 is suggested to result in uncontrolled activation of TGF-β/BMP signaling pathway, resulting in osteoproliferation (64). Yet another study using MSCs from the spinal entheseal sites of AS patients demonstrated that mineralization of MSCs was increased by the HLA-B*27–mediated spliced X-box–binding protein 1 (sXBP1)/retinoic acid receptor-β (RARB)/tissue-nonspecific alkaline phosphatase (TNAP) axis (65). Further elucidating the relationship between HLA-B*27 and the ALK2 and TNAP axes will be important for a better understanding the pathogenesis of new bone formation in SpA.

# ROLE OF HLA-B*27 IN SpA PATHOGENESIS

HLA-I molecules are highly polymorphic, comprising a heavy chain, a light chain [β2-microglobulin (β2m)] and a short peptide. The connections between HLA-B*27 and SpA must correlate with their distinct properties from the other HLA-I allotypes. HLA-B*27 allotypes have a preference for peptides with an arginine at the P2 position, as do some other allotypes. A genetic study argues that the strongest association with SpA was a unique asparagine at residue 97 in HLA-B*27, which lies on the floor of the peptide-binding groove and determines the specificity for C-terminal residue of the peptide (66). In addition, quantitative repertoire differences at the peptide C-terminus were also identified between SpA-linked and

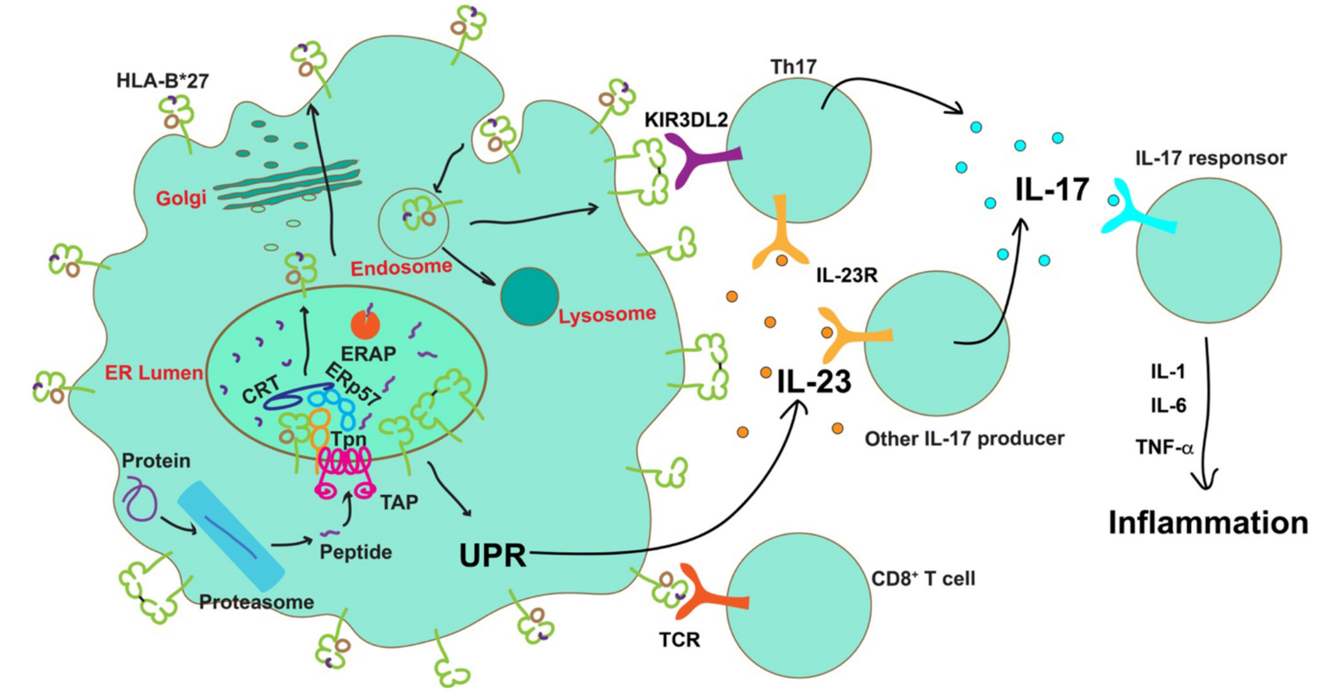


**FIGURE 2 |** Possible roles of HLA-B*27 in the pathogenesis of AS. After transport into the ER by the transporter associated with antigen processing (TAP), peptides are assembled onto nascent MHC class I molecules by the peptide loading complex (PLC), comprising TAP, tapasin (Tpn) calreticulin (CRT) and ERp57, and further trimmed by ERAP. The presence of certain ERAP haplotypes results in a significantly decreased number of peptides optimal for HLA-B*27, leading to accumulation of misfolded proteins in the ER or expression of sub-optimally or neoantigen-loaded HLA-B*27 on the cell surface. Neoantigen loaded HLA-B*27 on the cell surface could induce activation of CD8$^+$ T cells. Protein misfolding in the ER can cause ER stress and activation of the unfolded protein response (UPR) to induce IL-23 secretion. IL-23 further activates the IL-23/IL-17 axis. On the other hand, free heavy chains or disulfide bond-linked homodimers of HLA-B*27 can be formed and expressed on the cell surface during HLA-B*27 recycling via the endocytic pathway. Engagement of these aberrant species by KIR3DL2 on the surface of Th17 cells is known to enhance their survival, proliferation, and IL-17 expression. IL-17 can promote the release of proinflammatory cytokines to establish inflammation.

non-SpA-linked HLA-B*27 allotypes (67). If peptide specificity is the basis for connection between HLA-B*27 and SpA, C-terminal residues might be critical. On the other hand, different from many other HLA-I allotypes, HLA-B*27 allotypes have a free cysteine at the residue 67, which was shown to promote the formation of misfolded versions, including free heavy chains (FHC) and intermolecular disulfide bond-mediated heavy chain homodimers (B*$27_2$) (68). It was reported that the non-SpA-linked allotype HLA-B*27:09 show reduced HLA-B*27 dimer formation, compared with SpA-associated HLA-B*27 allotype HLA-B*27:05 (69). However, a more recent study examining the abilities of eight HLA-B*27 allotypes to form homodimers suggests that homodimer formation does not reflect different associations with SpA (70). In line with these features of HLA-B*27, there are three main hypotheses of the role played by HLA-B*27 in SpA pathogenesis, as illustrated in **Figure 2**.

## Arthritogenic Peptide Theory: HLA–B*27 Presents Specific Antigens to CD8$^+$ T Cells

Given the strong associations between SpA and HLA-B*27 and the major role of HLA-I in antigen presentation to CD8$^+$ T cells, SpA was initially considered as a disorder of adaptive immunity. The "arthritogenic peptide theory" proposed that, under certain conditions, some HLA-B*27 allotypes present self-peptides to specific CD8$^+$ T cells, causing damage to self-tissues. An early study showed that higher levels of CD8$^+$ T cells in the entheses of SpA patients (71). A recent study also showed that CD8$^+$ T cells are recruited to synovial fluid from the blood of AS patients (72). Additionally, self-peptide-responsive CD8$^+$ T cells have been identified in the peripheral blood mononuclear cells (PBMCs) or synovial fluid of some AS patients (73, 74). However, the arthritogenic peptide theory is challenged by the findings in transgenic rat models. Onset and severity of SpA manifestations were not affected in CD8α-deficient rat model (75), indicating that CD8$^+$ T cells are not essential for the development of SpA, nudging investigators to look beyond the classical functions of MHC-I molecules to better understand disease precipitation.

## HLA-B*27 Induces ER Stress

Based on recent models, SpA is suggested to be more related to innate rather than adaptive immune responses. As described above, specific properties of HLA-B*27 are known to cause the accumulation of unfolded and misfolded HLA-B*27 in the ER, which are suggested to potentially activate UPR. UPR is a set of intracellular pathways that signal the presence of ER stress upon sensing of misfolded proteins. Sustained over-activation of the UPR has been implicated in multiple inflammatory diseases. The possible relationship between HLA-B*27 misfolding, UPR and SpA pathogenesis was first investigated in an HLA-B*27 transgenic rat models. The rat line 33-3 expressing 55 copies of HLA-B*27:05 and 66 copies of hβ2m had predominant gut disease. High copies of B*27 in this model lead to protein misfolding in the ER resulting in UPR activation in bone marrow-derived macrophages and in the gut. In contrast, similar experiments with HLA-B*07, which is from a different supertype with different folding properties, did not show HLA-I misfolding and UPR activation (76). HLA-B*27 misfolding-activated UPR in bone marrow-derived macrophages augmented IL-23 expression induced by LPS. Increase of UPR in the colon of HLA-B*27 transgenic rats was associated with increased activation of the IL-23/IL-17 axis (39). These findings provide possible links between HLA-B*27 misfolding and dysregulation of immune responses in SpA patients. However, inconsistent with the HLA-B*27 misfolding-UPR theory, increasing β2m copy numbers (which promotes HLA-B*27 folding in the ER and mitigates UPR) enhanced the severity of AS symptoms in HLA-B*27 transgenic rats (77). These results argue against the crucial role of HLA-B*27 misfolding in the ER for SpA development.

Studies from human subjects are also inconclusive. While in one study, elevated ER stress and UPR activation was observed in monocytes/macrophages from peripheral blood and synovial fluid of AS patients (78), ER stress or UPR activation was not seen in other studies with peripheral blood monocyte-derived macrophages from HLA-B*27$^+$ AS patients (79, 80). UPR activation was also not observed in the gut of individuals with AS (40). Of note, HLA-B*27 has recently been shown to lower the threshold for UPR induction, and UPR might be activated in AS patients by a combination of HLA-B*27 misfolding and infection by certain bacteria (81). In addition, there appears to be a relationship between autophagy and IL-23 expression in the ileum of HLA-B*27$^+$ AS patients (40).

## Cell Surface HLA–B*27 FHC and B*27$_2$ Regulate Immune Responses

HLA-B*27 FHC or B*27$_2$ are expressed on the cell surface, and these aberrant forms of HLA-B*27 could be generated in the endosomes during the constitutive recycling of the cell surface HLA-B*27 molecules (82). These non-conventional HLA-B*27 conformers are suggested to be more readily detected in HLA-B*27$^+$ AS patients than HLA-B*27$^+$ healthy donors (83). Inhibiting HLA-B*27$_2$ with the monoclonal antibody HD6 prevented IL-17 secretion by PBMCs from HLA-B*27$^+$ SpA patients (83). These findings suggest cell surface aberrant HLA-B*27 conformers might be induced in SpA and play important roles in SpA pathogenesis.

Leukocyte immunoglobulin (Ig)-like receptors (LILR), such as LILRA1, LILRB2 (84), and LILRB5 (85) are shown to recognize HLA-B*27 FHC or B*27$_2$. B*27$_2$ also binds to the paired immunoglobulin-like receptors (PIR) in rodents (86), which share homology in their ligand binding domains to the LILR families in humans. How the interaction between B*27$_2$ and LILR/PIR connects to enhanced inflammation is not yet well-established. On the other hand, HLA-B*27 FHC and B*27$_2$ were reported to bind the killer cell immunoglobulin-like receptor (KIR)-KIR3DL2 more strongly than other ligands (87). Engagement of KIR3DL2 by B*27$_2$ promotes the survival of KIR3DL2$^+$ NK cells, and peripheral blood NK cells from SpA patients showed higher cytotoxicity than those from RA patients and healthy controls (88). In addition to NK cells, KIR3DL2 is also expressed on CD4$^+$ T cells. The engagement of KIR3DL2 by HLA-B*27 FHC and B*27$_2$ was shown to increase the survival and proliferation of Th17 cells from AS patients (89), consistent with the finding that Th17 cells are enriched in the peripheral blood and synovial fluid of patients with early axSpA (90). Importantly, if the interaction between aberrant HLA-B*27 conformers and a specific receptor is critical for AS pathogenesis, this receptor is likely to be conserved across species, given that HLA-B*27 induces AS or AS-like disease in both human and rodent animals.

## Role of ERAP

Although there is a strong association between HLA-B*27 and AS, a majority of the HLA-B*27 individuals never develop disease, implying that HLA-B*27 is not the only risk factor. In addition to cytokine pathways discussed previously, GWAS studies have shown that the ER aminopeptidase ERAP1 has strong genetic association with AS, in addition to HLA-B*27 (34). ERAP1 and ERAP2 are ER-localized aminopeptidases, which trim ER peptides at the N-terminus to produce peptides of optimal length for HLA-I binding and presentation (91). ERAP1 association is only significant in specific subsets of individuals who are HLA-B*27$^+$ (92) or HLA-B*40$^+$ (66), implying that AS development depends on the combined effect of HLA-B and ERAP1. The association between ERAP1 and HLA-B*27 has been widely studied. As an ER-localized aminopeptidase, ERAP1 influences both the quality and quantity of peptides available for HLA-B*27 in the ER, and thus, influences the peptide repertoire of HLA-B*27 expressed on the cell surface. ERAP1 single nucleotide polymorphisms (SNP), which are distinctly associated with AS, are proposed to cause differences in peptide cleavage efficiency and substrate selectivity, further emphasizing the importance of peptide loading of HLA-B*27 in AS pathogenesis. The most significant ERAP polymorphisms associated with AS include rs30187, that encodes the K528R variant, and rs27044, that encodes the Q730E variant (33, 93). In general, K528 is associated with more efficient cleavage, while Q730 is associated with increased preference for shorter peptides. However, the effect of each residue on peptide cleavage efficiency is not absolute, but rather is dependent on ERAP1 polymorphisms at other residues (94, 95). Early studies have shown that ERAP1 allotypes with high cleavage efficiency are risk factors for AS development, while allotypes with diminished activity are protective. Follow-up studies from Reeves et al. suggest that, based on their cleavage efficiencies, ERAP1 variants can be roughly divided into hyperactive (over-trimming), normal and hypoactive (inefficient trimming) allotypes. Hyperactive allotypes are always detrimental and hypoactive allotypes are detrimental when two are co-expressed, and these scenarios are uniquely present in AS patients (95). The study also found that

ERAP1 allotype combinations from AS patients are less efficient in promoting HLA-B*27:05 expression than those from non-AS controls, suggesting abnormal peptide cleavage (either too strong or too weak) and inefficient peptide loading of HLA-B*27 are responsible for AS onset (95). However, this theory was challenged by analysis of ERAP1 SNP haplotypes from a larger cohort of AS patients and RA controls, which argues that ERAP1 allele combinations in AS patients reported by Reeves et al. are pretty rare, and that most ERAP1 allele combinations from AS patients are shared with RA controls and do not necessarily contain one hyperactive allotype or two hypoactive allotypes (96). Clearly, further studies are needed to better understand ERAP variant prevalence and functions in SpA.

One possible consequence of the presence of risk ERAP1 allotypes is that they exclusively produce arthritogenic peptides, while the other possible consequence is that they induce HLA-B*27 misfolding and expression of aberrant forms of HLA-B*27, resulting from abnormal peptide trimming in the ER, which causes a shortage of optimal peptides for HLA-B*27 loading. The current data on the effect of ERAP1 polymorphisms on the cell surface expression of misfolded HLA-B*27 are still controversial (97–99). This is at least in part due to the complex effects of polymorphic residues on the function of ERAP1. Predictions of the cleavage efficiencies of ERAP1 allotypes based on individual residue occurrences, frequently adopted in the literature, may not be fully accurate, and the effects of altered cleavage efficiencies could be variable for different HLA-B allotypes. In addition, the combined effects of ERAP1 and ERAP2 should be considered. ERAP1 and ERAP2 can form heterodimers to trim peptides with enhanced efficiency (100). Polymorphisms of ERAP2 also affect its activity (91) and thus the genotype of ERAP2 is an important consideration while studying the effect of ERAP1 on HLA-B*27 misfolding.

## CONCLUDING REMARKS

HLA-B*27 associated with the development of articular and extra-articular manifestations of SpA. Presence of HLA-B*27 also drives structural damage in the axial skeletal of individuals with axSpA. These clinical associations have brought to focus the role of HLA-B*27 in the pathogenesis of SpA. Since HLA-I molecules are highly polymorphic, the unique properties of HLA-B*27 have drawn much attention. Accumulating evidence implicates HLA-B*27 molecules in disease through a number of mechanisms, including presenting arthritogenic peptides, misfolding and induction of UPR, and producing aberrant conformers on the cell surface that engage innate immune receptors. However, the primary causal factor is yet to be identified. Misfolding and aberrant structural variants of HLA-B27 are linked to the IL-23/IL-17 axis, which was shown to be critical from GWAS studies and trials of novel biologic therapies. Recent advances indicate that SpA is caused by a combination of HLA-B*27 and other genetic factors. It is important to recognize that, although the association with HLA-B*27 is strong, various forms of SpA also occur in individuals carrying other HLA-B alleles. Studies of the common characteristics of HLA-B*27 with such allotypes might shed further light on the pathogenesis of SpA.

## AUTHOR CONTRIBUTIONS

CK, JG, and MR designed the review structure. CK and JG wrote and edited the manuscript. MR, SB, and VN edited the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health, United States (Grants R01AI044115, and R21AI126054 to MR).

## ACKNOWLEDGMENTS

We thank the many contributors to the field and all present and previous members of the Raghavan lab for their contributions to HLA class I projects.

## REFERENCES

1. Rudwaleit M. New approaches to diagnosis and classification of axial and peripheral spondyloarthritis. *Curr Opin Rheumatol.* (2010) 22:375–80. doi: 10.1097/BOR.0b013e32833ac5cc
2. Stolwijk C, van Onna M, Boonen A, van Tubergen A. Global prevalence of spondyloarthritis: a systematic review and meta-regression analysis. *Arthritis Care Res.* (2016) 68:1320–31. doi: 10.1002/acr.22831
3. Schmitt SK. Reactive arthritis. *Infect Dis Clin North Am.* (2017) 31:265–77. doi: 10.1016/j.idc.2017.01.002
4. Chi CC, Tung TH, Wang J, Lin YS, Chen YF, Hsu TK, et al. Risk of uveitis among people with psoriasis: a nationwide cohort study. *JAMA Ophthalmol.* (2017) 135:415–22. doi: 10.1001/jamaophthalmol.2017.0569
5. Smith JA. Update on ankylosing spondylitis: current concepts in pathogenesis. *Curr Allergy Asthma Rep.* (2015) 15:489. doi: 10.1007/s11882-014-0489-6
6. De Vos M, Mielants H, Cuvelier C, Elewaut A, Veys E. Long-term evolution of gut inflammation in patients with spondyloarthropathy. *Gastroenterology.* (1996) 110:1696–703. doi: 10.1053/gast.1996.v110.pm8964393
7. Karreman MC, Luime JJ, Hazes JMW, Weel A. The prevalence and incidence of axial and peripheral spondyloarthritis in inflammatory bowel disease: a systematic review and meta-analysis. *J Crohns Colitis.* (2017) 11:631–42. doi: 10.1093/ecco-jcc/jjw199
8. Ward MM, Deodhar A, Gensler LS, Dubreuil M, Yu D, Khan MA, et al. 2019 update of the American College of Rheumatology/Spondylitis Association of America/Spondyloarthritis Research and treatment network recommendations for the treatment of ankylosing spondylitis and nonradiographic axial spondyloarthritis. *Arthritis Care Res.* (2019) 71:1285–99. doi: 10.1002/acr.24025
9. Hersh AH, Stecher RM, Solomon WM, Wolpaw R, Hauser H. Heredity in ankylosing spondylitis. *Am J Hum Genet.* (1950) 2:391–408.
10. Rudwaleit M, Haibel H, Baraliakos X, Listing J, Marker-Hermann E, Zeidler H, et al. The early disease stage in axial spondylarthritis: results from

the German spondyloarthritis inception cohort. *Arthritis Rheum.* (2009) 60:717–27. doi: 10.1002/art.24483

11. Arevalo M, Gratacos Masmitja J, Moreno M, Calvet J, Orellana C, Ruiz D, et al. Influence of HLA-B27 on the ankylosing spondylitis phenotype: results from the REGISPONSER database. *Arthritis Res Ther.* (2018) 20:221. doi: 10.1186/s13075-018-1724-7
12. Lin H, Gong YZ. Association of HLA-B27 with ankylosing spondylitis and clinical features of the HLA-B27-associated ankylosing spondylitis: a meta-analysis. *Rheumatol Int.* (2017) 37:1267–80. doi: 10.1007/s00296-017-3741-2
13. Chung HY, Machado P, van der Heijde D, D'Agostino MA, Dougados M. HLA-B27 positive patients differ from HLA-B27 negative patients in clinical presentation and imaging: results from the DESIR cohort of patients with recent onset axial spondyloarthritis. *Ann Rheum Dis.* (2011) 70:1930–6. doi: 10.1136/ard.2011.152975
14. Maksymowych WP, Wichuk S, Dougados M, Jones H, Szumski A, Bukowski JF, et al. MRI evidence of structural changes in the sacroiliac joints of patients with non-radiographic axial spondyloarthritis even in the absence of MRI inflammation. *Arthritis Res Ther.* (2017) 19:126. doi: 10.1186/s13075-017-1342-9
15. Dougados M, Sepriano A, Molto A, van Lunteren M, Ramiro S, de Hooge M, et al. Sacroiliac radiographic progression in recent onset axial spondyloarthritis: the 5-year data of the DESIR cohort. *Ann Rheum Dis.* (2017) 76:1823–8. doi: 10.1136/annrheumdis-2017-211596
16. Coates LC, Baraliakos X, Blanco FJ, Blanco-Morales E, Braun J, Chandran V, et al. The phenotype of axial spondyloarthritis: is it dependent on HLA-B27 status? *Arthritis Care Res.* (2020) doi: 10.1002/acr.24174. [Epub ahead of print].
17. Chandran V. *Genetic Determinants of Psoriatic Arthritis* (PhD thesis) (2013).
18. Queiro R, Sarasqueta C, Belzunegui J, Gonzalez C, Figueroa M, Torre-Alonso JC. Psoriatic spondyloarthropathy: a comparative study between HLA-B27 positive and HLA-B27 negative disease. *Semin Arthritis Rheum.* (2002) 31:413–8. doi: 10.1053/sarh.2002.33470
19. Salvarani C, Fries W. Clinical features and epidemiology of spondyloarthritides associated with inflammatory bowel disease. *World J Gastroenterol.* (2009) 15:2449–55. doi: 10.3748/wjg.15.2449
20. Kavadichanda CG, Seth G, Kumar G, Gulati R, Negi VS. Clinical correlates of HLA-B*27 and its subtypes in enthesitis-related arthritis variant of juvenile idiopathic arthritis in south Indian Tamil patients. *Int J Rheum Dis.* (2019) 22:1289–96. doi: 10.1111/1756-185X.13551
21. Adrovic A, Barut K, Sahin S, Kasapcopur O. Juvenile spondyloarthropathies. *Curr Rheumatol Rep.* (2016) 18:55. doi: 10.1007/s11926-016-0603-y
22. Watad A, Cuthbert RJ, Amital H, McGonagle D. Enthesitis: much more than focal insertion point inflammation. *Curr Rheumatol Rep.* (2018) 20:41. doi: 10.1007/s11926-018-0751-3
23. Gracey E, Dumas E, Yerushalmi M, Qaiyum Z, Inman RD, Elewaut D. The ties that bind: skin, gut and spondyloarthritis. *Curr Opin Rheumatol.* (2019) 31:62–9. doi: 10.1097/BOR.0000000000000569
24. Van Praet L, Van den Bosch FE, Jacques P, Carron P, Jans L, Colman R, et al. Microscopic gut inflammation in axial spondyloarthritis: a multiparametric predictive model. *Ann Rheum Dis.* (2013) 72:414–7. doi: 10.1136/annrheumdis-2012-202135
25. Taurog JD, Richardson JA, Croft JT, Simmons WA, Zhou M, Fernandez-Sueiro JL, et al. The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. *J Exp Med.* (1994) 180:2359–64. doi: 10.1084/jem.180.6.2359
26. Lin P, Bach M, Asquith M, Lee AY, Akileswaran L, Stauffer P, et al. HLA-B27 and human beta2-microglobulin affect the gut microbiota of transgenic rats. *PLoS ONE.* (2014) 9:e105684. doi: 10.1371/journal.pone.0105684
27. Costello ME, Ciccia F, Willner D, Warrington N, Robinson PC, Gardiner B, et al. Brief report: intestinal dysbiosis in ankylosing spondylitis. *Arthritis Rheumatol.* (2015) 67:686–91. doi: 10.1002/art.38967
28. Tito RY, Cypers H, Joossens M, Varkas G, Van Praet L, Glorieus E, et al. Brief report: dialister as a microbial marker of disease activity in spondyloarthritis. *Arthritis Rheumatol.* (2017) 69:114–21. doi: 10.1002/art.39802
29. Breban M, Tap J, Leboime A, Said-Nahal R, Langella P, Chiocchia G, et al. Faecal microbiota study reveals specific dysbiosis in spondyloarthritis. *Ann Rheum Dis.* (2017) 76:1614–22. doi: 10.1136/annrheumdis-2016-211064
30. Ansalone C, Utriainen L, Milling S, Goodyear CS. Role of gut inflammation in altering the monocyte compartment and its osteoclastogenic potential in HLA-B27-transgenic rats. *Arthritis Rheumatol.* (2017) 69:1807–15. doi: 10.1002/art.40154
31. Asquith M, Davin S, Stauffer P, Michell C, Janowitz C, Lin P, et al. Intestinal metabolites are profoundly altered in the context of HLA-B27 expression and functionally modulate disease in a rat model of spondyloarthritis. *Arthritis Rheumatol.* (2017) 69:1984–95. doi: 10.1002/art.40183
32. Ciccia F, Guggino G, Rizzo A, Alessandro R, Luchetti MM, Milling S, et al. Dysbiosis and zonulin upregulation alter gut epithelial and vascular barriers in patients with ankylosing spondylitis. *Ann Rheum Dis.* (2017) 76:1123–32. doi: 10.1136/annrheumdis-2016-210000
33. Wellcome Trust Case Control Consortium, Australo-Anglo-American Spondylitis Consortium, Burton PR, Clayton DG, Cardon LR, Craddock N, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet.* (2007) 39:1329–37. doi: 10.1038/ng.2007.17
34. Australo-Anglo-American Spondyloarthritis Consortium, Reveille JD, Sims AM, Danoy P, Evans DM, Leo P, et al. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat Genet.* (2010) 42:123–7. doi: 10.1038/ng.513
35. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet.* (2016) 48:510–8. doi: 10.1038/ng.3528
36. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science.* (2006) 314:1461–3. doi: 10.1126/science.1135245
37. Ciccia F, Bombardieri M, Principato A, Giardina A, Tripodo C, Porcasi R, et al. Overexpression of interleukin-23, but not interleukin-17, as an immunologic signature of subclinical intestinal inflammation in ankylosing spondylitis. *Arthritis Rheum.* (2009) 60:955–65. doi: 10.1002/art.24389
38. Wang X, Lin Z, Wei Q, Jiang Y, Gu J. Expression of IL-23 and IL-17 and effect of IL-23 on IL-17 production in ankylosing spondylitis. *Rheumatol Int.* (2009) 29:1343–7. doi: 10.1007/s00296-009-0883-x
39. DeLay ML, Turner MJ, Klenk EI, Smith JA, Sowders DP, Colbert RA. HLA-B27 misfolding and the unfolded protein response augment interleukin-23 production and are associated with Th17 activation in transgenic rats. *Arthritis Rheum.* (2009) 60:2633–43. doi: 10.1002/art.24763
40. Ciccia F, Accardo-Palumbo A, Rizzo A, Guggino G, Raimondo S, Giardina A, et al. Evidence that autophagy, but not the unfolded protein response, regulates the expression of IL-23 in the gut of patients with ankylosing spondylitis and subclinical gut inflammation. *Ann Rheum Dis.* (2014) 73:1566–74. doi: 10.1136/annrheumdis-2012-202925
41. Raychaudhuri SP, Raychaudhuri SK. IL-23/IL-17 axis in spondyloarthritis-bench to bedside. *Clin Rheumatol.* (2016) 35:1437–41. doi: 10.1007/s10067-016-3263-4
42. Kontoyiannis D, Pasparakis M, Pizarro TT, Cominelli F, Kollias G. Impaired on/off regulation of TNF biosynthesis in mice lacking TNF AU-rich elements: implications for joint and gut-associated immunopathologies. *Immunity.* (1999) 10:387–98. doi: 10.1016/S1074-7613(00)80038-2
43. Jacques P, Lambrecht S, Verheugen E, Pauwels E, Kollias G, Armaka M, et al. Proof of concept: enthesitis and new bone formation in spondyloarthritis are driven by mechanical strain and stromal cells. *Ann Rheum Dis.* (2014) 73:437–45. doi: 10.1136/annrheumdis-2013-203643
44. McGonagle D, Stockwin L, Isaacs J, Emery P. An enthesitis based model for the pathogenesis of spondyloarthropathy. additive effects of microbial adjuvant and biomechanical factors at disease sites. *J Rheumatol.* (2001) 28:2155–9.
45. Sherlock JP, Joyce-Shaikh B, Turner SP, Chao CC, Sathe M, Grein J, et al. IL-23 induces spondyloarthropathy by acting on ROR-$\gamma t^{+}$ $CD3^{+}CD4^{-}CD8^{-}$ entheseal resident T cells. *Nat Med.* (2012) 18:1069–76. doi: 10.1038/nm.2817
46. Deodhar A, Poddubnyy D, Pacheco-Tena C, Salvarani C, Lespessailles E, Rahman P, et al. Efficacy and safety of ixekizumab in the treatment of radiographic axial spondyloarthritis: sixteen-week results from a phase III randomized, double-blind, placebo-controlled trial in patients with prior inadequate response to or intolerance of tumor necrosis factor inhibitors. *Arthritis Rheumatol.* (2019) 71:599–611. doi: 10.1002/art.40753

47. Braun J, Baraliakos X, Deodhar A, Poddubnyy D, Emery P, Delicha EM, et al. Secukinumab shows sustained efficacy and low structural progression in ankylosing spondylitis: 4-year results from the MEASURE 1 study. *Rheumatology.* (2019) 58:859–68. doi: 10.1093/rheumatology/key375

48. Deodhar A, Gensler LS, Sieper J, Clark M, Calderon C, Wang Y, et al. Three multicenter, randomized, double-blind, placebo-controlled studies evaluating the efficacy and safety of ustekinumab in axial spondyloarthritis. *Arthritis Rheumatol.* (2019) 71:258–70. doi: 10.1002/art.40728

49. Baeten D, Ostergaard M, Wei JC, Sieper J, Jarvinen P, Tam LS, et al. Risankizumab, an IL-23 inhibitor, for ankylosing spondylitis: results of a randomised, double-blind, placebo-controlled, proof-of-concept, dose-finding phase 2 study. *Ann Rheum Dis.* (2018) 77:1295–302. doi: 10.1136/annrheumdis-2018-213328

50. Araujo EG, Englbrecht M, Hoepken S, Finzel S, Kampylafka E, Kleyer A, et al. Effects of ustekinumab versus tumor necrosis factor inhibition on enthesitis: results from the enthesial clearance in psoriatic arthritis (ECLIPSA) study. *Semin Arthritis Rheum.* (2019) 48:632–7. doi: 10.1016/j.semarthrit.2018.05.011

51. Cuthbert RJ, Fragkakis EM, Dunsmuir R, Li Z, Coles M, Marzo-Ortega H, et al. Brief report: group 3 innate lymphoid cells in human enthesis. *Arthritis Rheumatol.* (2017) 69:1816–22. doi: 10.1002/art.40150

52. Cuthbert RJ, Watad A, Fragkakis EM, Dunsmuir R, Loughenbury P, Khan A, et al. Evidence that tissue resident human enthesis gammadeltaT-cells can produce IL-17A independently of IL-23R transcript expression. *Ann Rheum Dis.* (2019) 78:1559–65. doi: 10.1136/annrheumdis-2019-215210

53. Reinhardt A, Yevsa T, Worbs T, Lienenklaus S, Sandrock I, Oberdorfer L, et al. Interleukin-23-dependent gamma/delta T Cells produce interleukin-17 and accumulate in the enthesis, aortic valve, and ciliary body in mice. *Arthritis Rheumatol.* (2016) 68:2476–86. doi: 10.1002/art.39732

54. Kenna TJ, Davidson SI, Duan R, Bradbury LA, McFarlane J, Smith M, et al. Enrichment of circulating interleukin-17-secreting interleukin-23 receptor-positive gamma/delta T cells in patients with active ankylosing spondylitis. *Arthritis Rheum.* (2012) 64:1420–9. doi: 10.1002/art.33507

55. Lories RJ, Matthys P, de Vlam K, Derese I, Luyten FP. Ankylosing enthesitis, dactylitis, and onychoperiostitis in male DBA/1 mice: a model of psoriatic arthritis. *Ann Rheum Dis.* (2004) 63:595–8. doi: 10.1136/ard.2003.013599

56. Capkova J, Ivanyi P. H-2 influence on ankylosing enthesopathy of the ankle (ANKENT). *Folia Biol.* (1992) 38:258–62.

57. van der Heijde D, Landewe R. Inhibition of spinal bone formation in AS: 10 years after comparing adalimumab to OASIS. *Arthritis Res Ther.* (2019) 21:225. doi: 10.1186/s13075-019-2045-1

58. McGonagle D, Marzo-Ortega H, O'Connor P, Gibbon W, Pease C, Reece R, et al. The role of biomechanical factors and HLA-B27 in magnetic resonance imaging-determined bone changes in plantar fascia enthesopathy. *Arthritis Rheum.* (2002) 46:489–93. doi: 10.1002/art.10125

59. Lories RJ, Luyten FP, de Vlam K. Progress in spondylarthritis. Mechanisms of new bone formation in spondyloarthritis. *Arthritis Res Ther.* (2009) 11:221. doi: 10.1186/ar2642

60. Lories RJ, Derese I, Luyten FP. Modulation of bone morphogenetic protein signaling inhibits the onset and progression of ankylosing enthesitis. *J Clin Invest.* (2005) 115:1571–9. doi: 10.1172/JCI23738

61. Lories RJ, Derese I, Ceuppens JL, Luyten FP. Bone morphogenetic proteins 2 and 6, expressed in arthritic synovium, are regulated by proinflammatory cytokines and differentially modulate fibroblast-like synoviocyte apoptosis. *Arthritis Rheum.* (2003) 48:2807–18. doi: 10.1002/art.11389

62. Xie Z, Wang P, Li Y, Deng W, Zhang X, Su H, et al. Imbalance between bone morphogenetic protein 2 and Noggin induces abnormal osteogenic differentiation of mesenchymal stem cells in ankylosing spondylitis. *Arthritis Rheumatol.* (2016) 68:430–40. doi: 10.1002/art.39433

63. El-Zayadi AA, Jones EA, Churchman SM, Baboolal TG, Cuthbert RJ, El-Jawhari JJ, et al. Interleukin-22 drives the proliferation, migration and osteogenic differentiation of mesenchymal stem cells: a novel cytokine that could contribute to new bone formation in spondyloarthropathies. *Rheumatology.* (2017) 56:488–93. doi: 10.1093/rheumatology/kew384

64. Grandon B, Rincheval-Arnold A, Jah N, Corsi JM, Araujo LM, Glatigny S, et al. HLA-B27 alters BMP/TGFβ signalling in Drosophila, revealing putative pathogenic mechanism for spondyloarthritis. *Ann Rheum Dis.* (2019) 78:1653–62. doi: 10.1136/annrheumdis-2019-215832

65. Liu CH, Raj S, Chen CH, Hung KH, Chou CT, Chen IH, et al. HLA-B27-mediated activation of TNAP phosphatase promotes pathogenic syndesmophyte formation in ankylosing spondylitis. *J Clin Invest.* (2019) 129:5357–73. doi: 10.1172/JCI125212

66. Cortes A, Pulit SL, Leo PJ, Pointon JJ, Robinson PC, Weisman MH, et al. Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat Commun.* (2015) 6:7146. doi: 10.1038/ncomms8146

67. Schittenhelm RB, Sian TC, Wilmann PG, Dudek NL, Purcell AW. Revisiting the arthritogenic peptide theory: quantitative not qualitative changes in the peptide repertoire of HLA-B27 allotypes. *Arthritis Rheumatol.* (2015) 67:702–13. doi: 10.1002/art.38963

68. Bowness P. Hla-B27. *Annu Rev Immunol.* (2015) 33:29–48. doi: 10.1146/annurev-immunol-032414-112110

69. Cauli A, Shaw J, Giles J, Hatano H, Rysnik O, Payeli S, et al. The arthritis-associated HLA-B$^{*}$27:05 allele forms more cell surface B27 dimer and free heavy chain ligands for KIR3DL2 than HLA-B$^{*}$27:09. *Rheumatology.* (2013) 52:1952–62. doi: 10.1093/rheumatology/ket219

70. Lim Kam Sian TCC, Indumathy S, Halim H, Greule A, Cryle MJ, Bowness P, et al. Allelic association with ankylosing spondylitis fails to correlate with human leukocyte antigen B27 homodimer formation. *J Biol Chem.* (2019) 294:20185–95. doi: 10.1074/jbc.RA119.010257

71. Laloux L, Voisin MC, Allain J, Martin N, Kerboull L, Chevalier X, et al. Immunohistological study of entheses in spondyloarthropathies: comparison in rheumatoid arthritis and osteoarthritis. *Ann Rheum Dis.* (2001) 60:316–21. doi: 10.1136/ard.60.4.316

72. Gracey E, Yao Y, Qaiyum Z, Lim M, Tang M, Inman RD. Altered cytotoxicity profile of CD8+ T cells in ankylosing spondylitis. *Arthritis Rheumatol.* (2020) 72:428–34. doi: 10.1002/art.41129

73. Fiorillo MT, Maragno M, Butler R, Dupuis ML, Sorrentino R. CD8$^{+}$ T-cell autoreactivity to an HLA-B27-restricted self-epitope correlates with ankylosing spondylitis. *J Clin Invest.* (2000) 106:47–53. doi: 10.1172/JCI9295

74. Atagunduz P, Appel H, Kuon W, Wu P, Thiel A, Kloetzel PM, et al. HLA-B27-restricted CD8+ T cell response to cartilage-derived self peptides in ankylosing spondylitis. *Arthritis Rheum.* (2005) 52:892–901. doi: 10.1002/art.20948

75. Taurog JD, Dorris ML, Satumtira N, Tran TM, Sharma R, Dressel R, et al. Spondylarthritis in HLA-B27/human $\beta_2$-microglobulin-transgenic rats is not prevented by lack of CD8. *Arthritis Rheum.* (2009) 60:1977–84. doi: 10.1002/art.24599

76. Turner MJ, Sowders DP, DeLay ML, Mohapatra R, Bai S, Smith JA, et al. HLA-B27 misfolding in transgenic rats is associated with activation of the unfolded protein response. *J Immunol.* (2005) 175:2438–48. doi: 10.4049/jimmunol.175.4.2438

77. Tran TM, Dorris ML, Satumtira N, Richardson JA, Hammer RE, Shang J, et al. Additional human β2-microglobulin curbs HLA-B27 misfolding and promotes arthritis and spondylitis without colitis in male HLA-B27-transgenic rats. *Arthritis Rheum.* (2006) 54:1317–27. doi: 10.1002/art.21740

78. Feng Y, Ding J, Fan CM, Zhu P. Interferon-γ contributes to HLA-B27-associated unfolded protein response in spondyloarthropathies. *J Rheumatol.* (2012) 39:574–82. doi: 10.3899/jrheum.101257

79. Zeng L, Lindstrom MJ, Smith JA. Ankylosing spondylitis macrophage production of higher levels of interleukin-23 in response to lipopolysaccharide without induction of a significant unfolded protein response. *Arthritis Rheum.* (2011) 63:3807–17. doi: 10.1002/art.30593

80. Ambarus CA, Yeremenko N, Baeten DL. Altered cytokine expression by macrophages from HLA-B27-positive spondyloarthritis patients without evidence of endoplasmic reticulum stress. *Rheumatol Adv Pract.* (2018) 2:rky014. doi: 10.1093/rap/rky014

81. Antoniou AN, Lenart I, Kriston-Vizi J, Iwawaki T, Turmaine M, McHugh K, et al. Salmonella exploits HLA-B27 and host unfolded protein responses to promote intracellular replication. *Ann Rheum Dis.* (2019) 78:74–82. doi: 10.1136/annrheumdis-2018-213532

82. Bird LA, Peh CA, Kollnberger S, Elliott T, McMichael AJ, Bowness P. Lymphoblastoid cells express HLA-B27 homodimers both intracellularly and at the cell surface following endosomal recycling. *Eur J Immunol.* (2003) 33:748–59. doi: 10.1002/eji.200323678

83. Payeli SK, Kollnberger S, Marroquin Belaunzaran O, Thiel M, McHugh K, Giles J, et al. Inhibiting HLA-B27 homodimer-driven immune cell inflammation in spondylarthritis. *Arthritis Rheum.* (2012) 64:3139–49. doi: 10.1002/art.34538
84. Allen RL, Raine T, Haude A, Trowsdale J, Wilson MJ. Leukocyte receptor complex-encoded immunomodulatory receptors show differing specificity for alternative HLA-B27 structures. *J Immunol.* (2001) 167:5543–7. doi: 10.4049/jimmunol.167.10.5543
85. Zhang Z, Hatano H, Shaw J, Olde Nordkamp M, Jiang G, Li D, et al. The leukocyte immunoglobulin-like receptor family member LILRB5 binds to HLA-class I heavy chains. *PLoS ONE.* (2015) 10:e0129063. doi: 10.1371/journal.pone.0129063
86. Kollnberger S, Bird LA, Roddis M, Hacquard-Bouder C, Kubagawa H, Bodmer HC, et al. HLA-B27 heavy chain homodimers are expressed in HLA-B27 transgenic rodent models of spondyloarthritis and are ligands for paired Ig-like receptors. *J Immunol.* (2004) 173:1699–710. doi: 10.4049/jimmunol.173.3.1699
87. Kollnberger S, Bird L, Sun MY, Retiere C, Braud VM, McMichael A, et al. Cell-surface expression and immune receptor recognition of HLA-B27 homodimers. *Arthritis Rheum.* (2002) 46:2972–82. doi: 10.1002/art.10605
88. Chan AT, Kollnberger SD, Wedderburn LR, Bowness P. Expansion and enhanced survival of natural killer cells expressing the killer immunoglobulin-like receptor KIR3DL2 in spondylarthritis. *Arthritis Rheum.* (2005) 52:3586–95. doi: 10.1002/art.21395
89. Wong-Baeza I, Ridley A, Shaw J, Hatano H, Rysnik O, McHugh K, et al. KIR3DL2 binds to HLA-B27 dimers and free H chains more strongly than other HLA class I and promotes the expansion of T cells in ankylosing spondylitis. *J Immunol.* (2013) 190:3216–24. doi: 10.4049/jimmunol.1202926
90. Shen H, Goodall JC, Gaston JS. Frequency and phenotype of T helper 17 cells in peripheral blood and synovial fluid of patients with reactive arthritis. *J Rheumatol.* (2010) 37:2096–9. doi: 10.3899/jrheum.100146
91. Lopez de Castro JA. How ERAP1 and ERAP2 shape the peptidomes of disease-associated MHC-I proteins. *Front Immunol.* (2018) 9:2463. doi: 10.3389/fimmu.2018.02463
92. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet.* (2011) 43:761–7. doi: 10.1038/ng.873
93. Harvey D, Pointon JJ, Evans DM, Karaderi T, Farrar C, Appleton LH, et al. Investigating the genetic association between ERAP1 and ankylosing spondylitis. *Hum Mol Genet.* (2009) 18:4204–12. doi: 10.1093/hmg/ddp371
94. Martin-Esteban A, Gomez-Molina P, Sanz-Bravo A, Lopez de Castro JA. Combined effects of ankylosing spondylitis-associated ERAP1 polymorphisms outside the catalytic and peptide-binding sites on the processing of natural HLA-B27 ligands. *J Biol Chem.* (2014) 289:3978–90. doi: 10.1074/jbc.M113.529610
95. Reeves E, Colebatch-Bourn A, Elliott T, Edwards CJ, James E. Functionally distinct ERAP1 allotype combinations distinguish individuals with ankylosing spondylitis. *Proc Natl Acad Sci USA.* (2014) 111:17594–9. doi: 10.1073/pnas.1408882111
96. Roberts AR, Appleton LH, Cortes A, Vecellio M, Lau J, Watts L, et al. ERAP1 association with ankylosing spondylitis is attributable to common genotypes rather than rare haplotype combinations. *Proc Natl Acad Sci USA.* (2017) 114:558–61. doi: 10.1073/pnas.1618856114
97. Haroon N, Tsui FW, Uchanska-Ziegler B, Ziegler A, Inman RD. Endoplasmic reticulum aminopeptidase 1 (ERAP1) exhibits functionally significant interaction with HLA-B27 and relates to subtype specificity in ankylosing spondylitis. *Ann Rheum Dis.* (2012) 71:589–95. doi: 10.1136/annrheumdis-2011-200347
98. Tran TM, Hong S, Edwan JH, Colbert RA. ERAP1 reduces accumulation of aberrant and disulfide-linked forms of HLA-B27 on the cell surface. *Mol Immunol.* (2016) 74:10–7. doi: 10.1016/j.molimm.2016.04.002
99. Chen L, Ridley A, Hammitzsch A, Al-Mossawi MH, Bunting H, Georgiadis D, et al. Silencing or inhibition of endoplasmic reticulum aminopeptidase 1 (ERAP1) suppresses free heavy chain expression and Th17 responses in ankylosing spondylitis. *Ann Rheum Dis.* (2016) 75:916–23. doi: 10.1136/annrheumdis-2014-206996
100. Evnouchidou I, Weimershaus M, Saveanu L, van Endert P. ERAP1-ERAP2 dimerization increases peptide-trimming efficiency. *J Immunol.* (2014) 193:901–8. doi: 10.4049/jimmunol.1302855

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership