



OPEN ACCESS

EDITED BY

Jerome Le Nours,
Monash University, Australia

REVIEWED BY

Robert David Miller,
University of New Mexico, United States
Karla Rodgers,
University of Oklahoma Health Sciences
Center, United States

*CORRESPONDENCE

Sofia Kossida

✉ sofia.kossida@igh.cnrs.fr

Véronique Giudicelli

✉ veronique.giudicelli@igh.cnrs.fr

RECEIVED 02 August 2024

ACCEPTED 19 September 2024

PUBLISHED 10 October 2024

CITATION

Debbagh C, Folch G, Jabado-Michaloud J, Giudicelli V and Kossida S (2024) Deciphering *Gorilla gorilla gorilla* immunoglobulin loci in multiple genome assemblies and enrichment of IMGT resources. *Front. Immunol.* 15:1475003. doi: 10.3389/fimmu.2024.1475003

COPYRIGHT

© 2024 Debbagh, Folch, Jabado-Michaloud, Giudicelli and Kossida. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Deciphering *Gorilla gorilla gorilla* immunoglobulin loci in multiple genome assemblies and enrichment of IMGT resources

Chahrazed Debbagh¹, Géraldine Folch¹,
Joumana Jabado-Michaloud¹, Véronique Giudicelli^{1*}
and Sofia Kossida^{1,2*}

¹The International ImMunoGeneTics® Information System (IMGT®), Institute of Human Genetics (IGH), National Center for Scientific Research (CNRS), University of Montpellier (UM), Montpellier, France,

²Institut Universitaire de France (IUF), Paris, France

Through the analysis of immunoglobulin genes at the IGH, IGK, and IGL loci from four *Gorilla gorilla gorilla* genome assemblies, IMGT® provides an in-depth overview of these loci and their individual variations in a species closely related to humans. The similarity between gorilla and human IG gene organization allowed the assignment of gorilla IG gene names based on their human counterparts. This study revealed significant findings, including variability in the IGH locus, the presence of known and new copy number variations (CNVs), and the accurate estimation of IGHG genes. The IGK locus displayed remarkable homogeneity and lacked the gene duplication seen in humans, while the IGL locus showed a previously unconfirmed CNV in the J-C cluster. The curated data from these analyses, available on the IMGT website, enhance our understanding of gorilla immunogenetics and provide valuable insights into primate evolution.

KEYWORDS

immunoglobulins, germline repertoire, immunogenetics, comparative genomics, adaptive immune response, *Gorilla gorilla gorilla*, IMGT

1 Introduction

Immunoglobulins (IG) and T-cell receptors (TR) are two types of antigen receptors that are responsible for the extraordinary specificity and memory for antigen recognition and binding, which characterize the adaptive immune response (1–3). Immunoglobulins consist of two types of chains—heavy chains (IGH) and light chains [Kappa (IGK) or Lambda (IGL)] (4)—which are encoded by four types of genes: variable (V), diversity (D), junction (J), and constant (C) (5, 6).

IG genes, which are distributed along the three IG loci—IGH, IGK, and IGL (7) (localized on three different chromosomes in humans and other vertebrates)—belong to

multigene families and are characterized by a high level of allelic polymorphism and great diversity, for example, for the D genes exclusively found in the IG heavy chain (4, 6, 8). Moreover, IG V, D, and J genes comprise specific motifs in their genomic sequences, such as recombination signals (RSs), which are responsible for generating the combinatorial diversity of the variable domains. Owing to their genetic complexity (1, 4), these genes are challenging to analyze and classify. Additionally, structural variations of the IG loci were shown between individuals of the same species.

IMGT^{®1}, the international ImMunoGeneTics[®] information system (9, 10), established in 1989, is a high-quality integrated knowledge resource that manages sequences from genome to proteome, and structural data for immunoglobulins and T-cell receptors in humans and other jawed vertebrates (11, 12). IMGT provides resources (database, tools, IMGT Repertoire as well as IG and TR genes and alleles reference sets) for jawed vertebrates for the analysis and understanding of immunogenetics.

Non-human primates are of great interest in comparative studies and biomedical research due to their close similarities to the human species (13). Several studies have explored the relationship between non-human primate evolution and human diseases, focusing particularly on segmental duplications found in great apes and humans, which are thought to play an important role in human susceptibility to diseases, such as the case of their impact on genes associated with Mendelian diseases (14, 15). Certain genes, alleles, and proteins could be implicated in causing diseases in humans, yet they may be associated with normal phenotypes in gorillas. For instance, this is observed in cases such as Moyamoya disease, causing deafness in humans (16), as well as instances of dementia and hypertrophic cardiomyopathy (17). However, these studies do not include an investigation of the adaptive immune system of gorillas, nor do they incorporate genomic data related to their immunoglobulins.

Indeed, in the study of the genetics of the adaptive immune system, IMGT has been engaged for over three decades in deciphering and characterizing IG and TR loci across various species of jawed vertebrates. A recent example is the rhesus monkey, a non-human primate, as documented in Reference (18). In our study, we utilized several genome assemblies of *Gorilla gorilla gorilla* (Western lowland gorilla) available in the National Center for Biotechnology Information (NCBI) repository, which were derived from different individuals and sequencing technologies, to establish the gene repertoire of the three IG loci. Four assemblies, corresponding to three individuals, were selected based on IMGT criteria.

Our analysis of IG loci in three *G. gorilla gorilla* (we will refer to “gorilla” by its abbreviated name throughout this article) individuals, aiming to identify significant similarities and some differences between humans and between gorilla individuals, provides insights into human evolutionary changes in their ability to fight infections and establish an appropriate immune system (19). Additionally, it offers detailed information on the evolutionary

origin of immunoglobulins and the phylogenetic relationships between primate species (17, 20). The gorilla is a protected species that cannot be utilized as an animal model directly. However, knowledge of its genome is critical in immunological disease research and treatment developments.

2 Methods

The annotation of the IG loci was performed according to the IMGT biocuration pipeline, as previously described (21). Gorilla IG genomic sequences were analyzed and annotated by comparison with the IMGT human reference sequences².

2.1 *G. gorilla gorilla* assembly selection

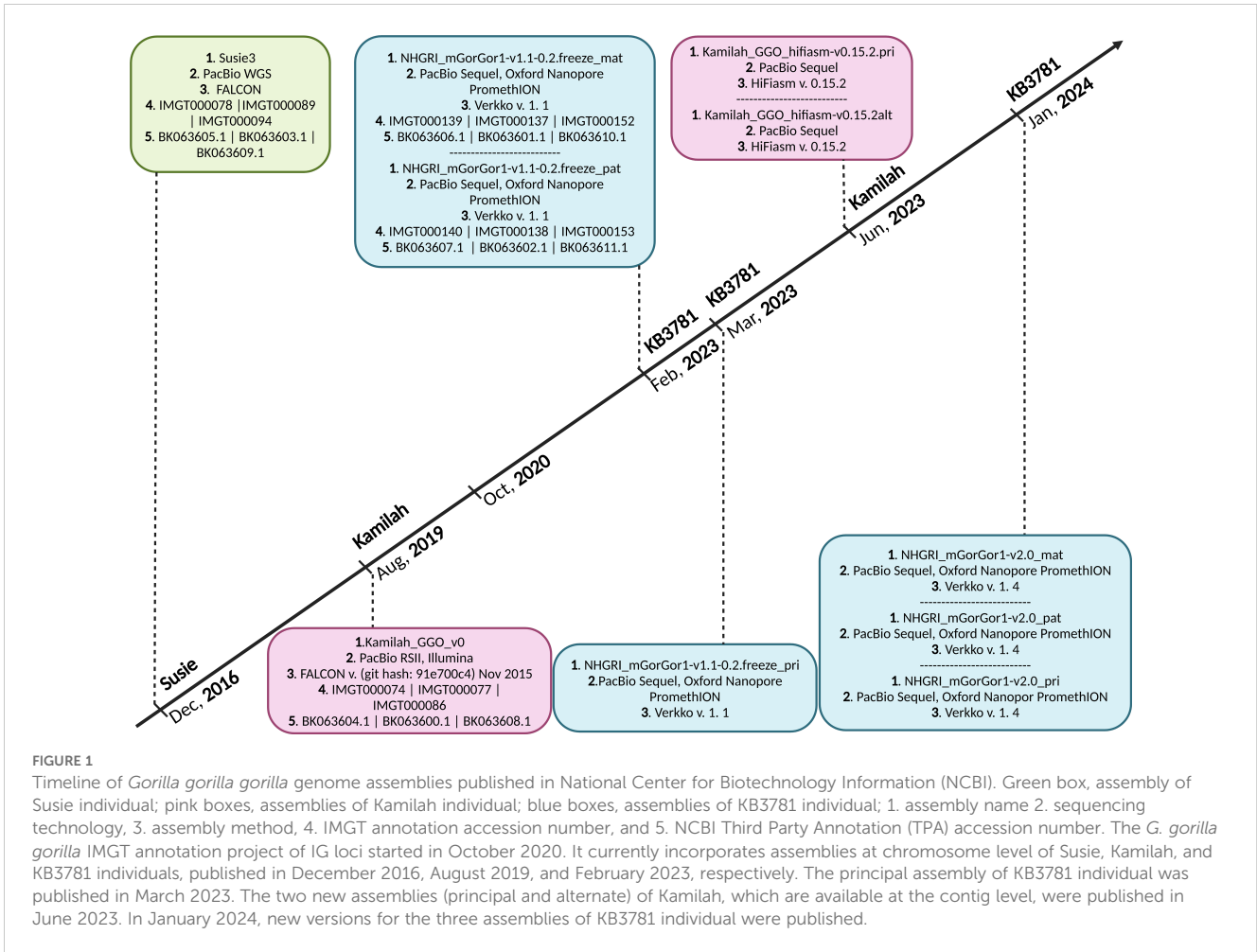
Seven assemblies of *Gorilla gorilla* species were available in October 2020 on the NCBI (22, 23); the Western lowland gorilla is the most sequenced genomic subspecies of gorillas. All of them were evaluated, and two—Kamilah_GGO_v0 (GenBank Assembly ID: GCA_008122165.1) (24), which was labeled “representative genome” of the *G. gorilla gorilla* at the NCBI, and Susie3 (GenBank Assembly ID: GCA_900006655.3) (25)—were chosen for the quality of their IG loci, which fulfilled the standard IMGT criteria for assembly selection³. In addition, two new *G. gorilla gorilla* assemblies of the same individual, KB3781, were added in the NCBI assembly [NCBI Datasets since 2024 (26)]: NHGRI_mGorGor1-v1.1-0.2.freeze_mat (GenBank Assembly ID: GCA_028885495.1), the maternal haplotype, and NHGRI_mGorGor1-v1.1-0.2.freeze_pat (GenBank Assembly ID: GCA_028885475.1), the paternal haplotype (27) (Figure 1). The analysis of these four assemblies, which are all characterized by a “Chromosome” assembly level [see the Glossary⁴ of the National Institutes of Health (NIH)], is incorporated in this study. In 2023, The NHGRI_mGorGor1-v1.1-0.2.freeze_pri assembly (GenBank Assembly ID: GCA_029281585.1), the principal haplotype of KB3781, became the “representative genome” of the *G. gorilla* species (it includes the maternal autosomes, unplaced sequence identified as maternal, chrX, chrY, and MT), and the three assemblies of the KB3781 genome were updated in 2024 (Figure 1). As for the two assemblies of the Kamilah individual published in 2023, Kamilah_GGO_hifiasm-v0.15.2.pri (GenBank Assembly ID: GCA_030174185.1) and Kamilah_GGO_hifiasm-v0.15.2.alt (GenBank Assembly ID: GCA_030174155.1) (28), the corresponding biocuration results are solely presented in the “Discussion” section for reasons that will be apparent further in this article.

2 <https://www.imgt.org/vquest/refseqh.html#refdir2>

3 <https://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGTassemblyselection.html>

4 <https://www.ncbi.nlm.nih.gov/datasets/docs/v2/glossary/>

1 <https://www.imgt.org/>



2.2 Locus sequence extraction from NCBI and integration in IMGT

For each assembly, the localization of the three IG loci (IGH, IGK, and IGL) on chromosomes was determined by comparison to the IMGT human IG reference set, using BLAST (29). The delimitations of the IGK and IGL loci were defined by the identification of the flanking non-IG genes, which are conserved among species upstream of the first IG gene and downstream of the last IG gene, called “IMGT bornes” (30). If and only if the distance of the “IMGT bornes” is over 10,000 bp from the first and last IG genes in 5’ and 3’, respectively, the delimitations of the IGK and IGL loci are defined by 10,000 bp upstream of the first IG gene and 10,000 bp downstream of the last IG gene. Due to the absence of “IMGT bornes” for IGH loci, the gorilla IGH loci were delimited by 10,000 bp upstream of the first IG gene and 11,000 bp (exclusively for gorilla IGH locus) downstream of the last IG gene. The corresponding nucleotide sequences were extracted from the NCBI chromosome sequences, and IMGT/LIGM-DB (31) entries were created.

5 <https://www.imgt.org/IMGTIndex/IMGTborne.php>

2.3 V, D, J, and C genes annotation

The V, D, J, and C genes were first detected and delimited along the IMGT/LIGM-DB (31) genomic sequences (IGH, IGK, and IGL loci), with IMGT/LIGMotif (32). IG genes were characterized and classified using alignments by BLAST (29) and Clustal Omega (33) and by implementing the IMGT unique numbering (34, 35) and annotation rules of the IMGT Scientific chart based on the IMGT-ONTOLOGY concepts (36) for the gene and allele functionalities⁶ and the setting of gorilla IG gene nomenclature⁷. Due to the extremely high sequence similarity between gorillas and humans (37), which was confirmed at the level of the IG loci in the early steps of biocuration (see section “Results” and Supplementary Tables 1-3), gorilla IG genes were named according to their human counterparts based on their sequence similarity and their position in the locus (nomenclature by orthology). Additional genes in gorilla loci compared to the

6 <https://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGTfunctionality.html>

7 <https://www.imgt.org/IMGTScientificChart/Nomenclature/IMGTnomenclature.html>

human species were named inserted genes. The incrementation of the number of the V gene sub-positions from 3' to 5', the number of the D gene sub-positions from 5' to 3', and the addition of the Latin alphabet letters, from 5' to 3' for the J and C genes. Another case of additional genes is the duplicated genes, which were named as the "initial" gene, with the addition of the letter "D". IG genes were integrated into IMGT/GENE-DB (38), and the synthesis of biocuration data regarding the loci, genes, alleles, and proteins into IMGT web resources for IG⁸.

The NCBI Third Party Annotation (TPA⁹) (39) accession numbers were provided for the three *G. gorilla gorilla* IG loci.

2.4 CNV characterization

The names of copy number variation (CNV) in gorilla IG loci are identical to those of human counterparts (30), if equivalent. For gorilla potential specific CNV, they are provisionally named CNVp, plus an incremental number.

3 Results

The three loci, IGH, IGK, and IGL, of genes were extracted from four *G. gorilla gorilla* genome assemblies of three individuals publicly available at the NCBI and annotated according to IMGT standards. Table 1 summarizes locus information and gives the total number of genes for each assembly, their accession numbers initially created in IMGT/LIGM-DB (31), and the ones assigned in the TPA database (39).

The resulting data for loci, assemblies, genes, alleles, sequences, proteins, expression cDNA, and statistics are available in IMGT web resources, and the detailed list is provided in Supplementary Table 4. The gorilla IG gene names were assigned by orthology with humans and according to IMGT gene nomenclature principles. The percentages of identity between the closest gorilla alleles and their human counterparts are reported in Supplementary Tables 1-3. The loci and genes data from Kamilah_GGO_v0 were chosen as a reference for the analysis of locus variations between gorilla individuals in terms of gene and allele contents. The IMGT gene order of V, D, J, C, and non-IG genes from 5' to 3' for each locus was initially established for the Kamilah_GGO_v0 assembly. According to this, the gene order of additional genes in the three other assemblies was identified (Supplementary Tables 1-3).

Figure 2 presents an overview of the number of annotated IG genes that are common or unique within the four gorilla assemblies. The locus IGH appears to be more variable in terms of common genes (73% of total annotated IGH genes) across the four assemblies compared to the IGK (94% of total common annotated IGK genes) and IGL (84% of total common annotated IGL genes) loci, which are more conserved among the three individuals. The heterogeneity

of the IGH locus between the four gorilla assemblies is particularly evident in the detected CNVs, as well as the duplication and insertion of genes throughout the locus (see section "*G. gorilla gorilla* IGH locus").

3.1 *G. gorilla gorilla* IGH locus

3.1.1 Localization and description of IGH locus

The IGH locus of the gorilla extends from 10 kb upstream of the most 5' gene in the locus, IGHV(III)-82, to 11 kb downstream of the most 3' gene, IGHA2. It comprises between 164 and 185 genes depending on the assembly, and all of them are in the sense orientation in the locus (Table 1). According to the description and annotation of the locus with "IMGT Labels¹⁰" (40), the IGH locus is composed of four clusters of the same gene type: 120 to 135 V genes (V-CLUSTER), 18 to 32 D genes (D-CLUSTER), 8 to 9 J genes (J-CLUSTER), and 4 to 13 C genes (C-CLUSTER). Its organization is very close to the human one. Interestingly, the eight known related proteins of the immune system (RPI) genes within the human IGH locus were also identified in gorilla assemblies (Table 1; Supplementary Figures 1-4; Supplementary Table 1).

3.1.2 IG gene organization in the gorilla IGH locus

3.1.2.1 IGHV gene cluster

Overall, 157 IGHV genes and 316 IGHV alleles were identified in the gorilla IGH loci of the four assemblies (Figure 2, Supplementary Table 1). Based on their high level of sequence similarity with human IGH genes, 105 gorilla V genes could be classified into eight subgroups and 52 others in three clans.

A phylogenetic tree was built from a sequence set including the first allele (the reference sequence of each gorilla gene) and the first allele for human genes. This phylogenetic tree was created to highlight the close similarity of genes between both species within a subgroup (Figure 3). The pseudogenes of the clan IGHV(III) intercalate with the subgroup IGHV3 because of the sequence similarity and according to the "IMGT IGH clan tree"¹¹. It shows that the gorilla subgroup/clan genes are grouped in the same branch with the corresponding human gene.

A total of 136 gorilla IGHV genes were named according to their human counterparts (Supplementary Table 1). The names of 29 additional IGHV genes, only present in the gorilla genome (underlined in green in Supplementary Table 1), were set by applying IMGT nomenclature for inserted genes or by taking into account the evidence of gene bloc duplication. Two duplicated blocs were identified, including genes IGHV3-41D to IGHV4-39D and genes IGHV(II)-62-1D to IGHV3-66D, which show over 99.5% of identity with the initial blocs IGHV3-41 to IGHV4-39 and IGHV(II)-62-1 to IGHV3-66, respectively.

8 <https://www.imgt.org/IMGTrepertoire/>

9 <https://www.ncbi.nlm.nih.gov/genbank/tpa/>

10 <https://www.imgt.org/IMGTIndex/labels.php>

11 <https://www.imgt.org/IMGTIndex/IGHVclans.php>

TABLE 1 Information about genome assembly and IGH, IGK, and IGL loci for the four *Gorilla gorilla gorilla* assemblies.

Taxon		<i>G. gorilla gorilla</i> (Western lowland gorilla), NCBI: TaxId: 9595			
Genome assembly		Kamilah_GGO_V0	Susie3	NHGRI_mGorGor1-v1.1-0.2.freeze_mat	NHGRI_mGorGor1-v1.1-0.2.freeze_pat
GenBank assembly ID		GCA_008122165.1	GCA_900006655.3	GCA_028885495.1	GCA_028885475.1
Sequencing technology/assembly method		PacBio RSII—Illumina/FALCON v. (git hash: 91e700c4) Nov 2015 method	PacBio WGS/FALCON	PacBio Sequel—Oxford Nanopore PromethION/Verkko V. 1.4	PacBio Sequel—Oxford Nanopore PromethION/Verkko V. 1.4
Isolate		Kamilah (stud number 0661)	NA	KB3781	KB3781
IGH locus	IMGT locus ID	Gorgor_IGH_1	Gorgor_IGH_2	Gorgor_IGH_3	Gorgor_IGH_4
	Chromosome	14	14	14	14
	GenBank chromosome sequence ID	CM017861.1	LT578329.1	CM054596.1	CM054572.1
	Locus positions	86452632-87593842, complement	86149321-87571187, complement	114386608-115697767, complement	118187703-119503099, complement
	IMGT locus orientation on the chromosome	REV	REV	REV	REV
	IMGT/LIGM-DB accession number	IMGT000074	IMGT000078	IMGT000139	IMGT000140
	TPA accession number	BK063604	BK063605	BK063606	BK063607
	Sequence length (bp)	1,141,211	1,421,867	1,311,160	1,315,397
	IG gene number	164	185	174	177
IGK locus	IMGT locus ID	Gorgor_IGK_1	Gorgor_IGK_2	Gorgor_IGK_3	Gorgor_IGK_4
	Chromosome	2A	2A	2A	2A
	GenBank chromosome sequence ID	CM017848.1	LT578338.1	CM054583.1	CM054559.1
	Locus positions	104237816-105130673, complement	17629982-18328204	94854218-96227846, complement	94027569-94921140, complement
	IMGT locus orientation on the chromosome	REV	FWD	REV	REV
	IMGT/LIGM-DB accession number	IMGT000077	IMGT000089	IMGT000137	IMGT000138
	TPA accession number	BK063600	BK063603	BK063601	BK063602
	Sequence length (bp)	892,858	698,223	1,373,629	893,572
	IG gene number	49	47	50	49
IGL locus	IMGT locus ID	Gorgor_IGL_1	Gorgor_IGL_2	Gorgor_IGL_3	Gorgor_IGL_4
	Chromosome	22	22	22	22
	GenBank chromosome sequence ID	CM017869.1	LT578337.1	CM054604.1	CM054580.1

(Continued)

TABLE 1 Continued

Taxon		<i>G. gorilla gorilla</i> (Western lowland gorilla), NCBI: TaxId: 9595			
Locus positions	4187005-5025689	5400563-6277394	16012935-16839685	14038805-14911038	
IMGT locus orientation on the chromosome	FWD	FWD	FWD	FWD	
IMGT/LIGM-DB accession number	IMGT000086	IMGT000094	IMGT000152	IMGT000153	
TPA accession number	BK063608	BK063609	BK063610	BK063611	
Sequence length (bp)	838,685	876,832	826,751	872,234	
IG gene number	90	100	93	102	

Based on this comparative approach, we also identified 29 human IGHV genes for which the gorilla counterparts cannot be found in the IGH locus in any of the four assemblies (underlined in yellow in **Supplementary Table 1**). Interestingly, all of them (except IGHV7-77) are located within well-known human CNVs (30).

Figure 4 shows the number of genes within subgroups and clans. Interestingly the majority of functional genes belong to the IGHV3 subgroup (**Supplementary Table 5**), as in humans. The human IGHV3 genes are known to be selected in response to superantigens (43, 44). The expansion of the subgroup IGHV3 could also be the result of the major CNV3 and the gorilla potential CNVp1. The detailed list of alleles per IGHV subgroup or clan and per functionality present in each assembly can be obtained from the section “Locus gene repertoire per IMGT annotated assembly¹²” of IMGT web resources (after selection of the species and the locus).

3.1.2.2 IGHD and IGJH gene clusters

The analysis of the gorilla IGH locus of four assemblies allowed the identification of 32 IGHD genes classified in seven sets as in humans and 43 IGHD alleles of which 29 are functional (**Supplementary Table 1**). Fifteen consecutive genes (including IGHD6-5-1, IGHD1-5-2, IGHD2-5-3, IGHD3-5-4, IGHD4-5-5, IGHD5-5-6, IGHD6-6, IGHD1-7, IGHD2-8, IGHD3-10, IGHD4-11, IGHD5-12, IGHD6-13, and IGHD1-14) are missing in Kamilah_GGO_V0 assembly due to a 20-kb gap within the D-CLUSTER area (**Supplementary Figure 1**). However, all IGHD genes were recovered in the new assemblies of Kamilah’s genome (data not shown, see section ‘Assemblies of “Kamilah” individual’). Among them, 26 have been named according to their human counterparts. Additionally, six new IGHD genes were identified (namely, GHD6-5-1, IGHD1-5-2, IGHD2-5-3, IGHD3-5-4, IGHD4-5-5, and IGHD5-5-6); they may be gorilla specific since human counterparts were not found in the 16 assemblies processed by IMGT. Based on the results obtained from the four assemblies, it appears that the gorilla IGH locus does not include an ortholog for the human IGHD3-9.

Nine IGJH genes and 16 alleles have been identified and localized in the gorilla IGH locus. The IGJH gene organization is comparable to the one of the human J-CLUSTER. They were classified into six subsets, and all of them were named according to their human counterparts. The gorilla IGJH2 gene was not found in the Kamilah_GGO_v0 assembly. It is also missing in the recent assembly version of this genome, Kamilah_GGO_hifiasm-v0.15.2.pri, but it is present in Kamilah_GGO_hifiasm-v0.15.2.alt (data not shown, see section ‘Assemblies of “Kamilah” individual’). Therefore, this variation could correspond to a CNV in the gorilla IGH locus.

3.1.2.3 IGHC gene cluster

A total of 13 gorilla IGHC genes and 29 alleles were identified, taking into account the four assemblies (**Supplementary Table 1**; **Supplementary Figures 1-4**; Dynamic gene tables per IMGT group

¹² https://www.imgt.org/IMGTrepertoire/LocusGenes/locusdesc/assembly_compare.php

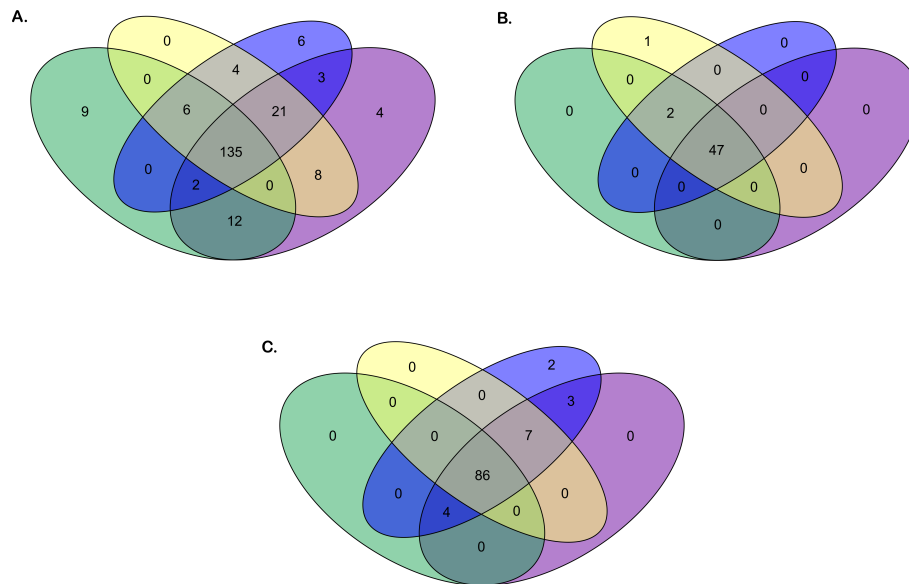


FIGURE 2

Venn diagrams of the number of genes for the four *Gorilla gorilla gorilla* assemblies. Green oval: Kamilah_GGO_v0 assembly. Purple oval: Susie3 assembly. Yellow oval: NHGRI_mGorGor1-v1.1-0.2.freeze_mat assembly. Blue oval: NHGRI_mGorGor1-v1.1-0.2.freeze_pat assembly. The Venn diagrams in panels (A–C) show respectively the common or unique IGH, IGK, and IGL genes [V, (D), J, and C] between the four *G. gorilla gorilla* assemblies.

and species¹³). These genes code for the five isotype classes: IgM (IGHM), IgD (IGHD), IgG (IGHG1, IGHG2, IGHG3A, IGHG3B, IGHG3C, IGHG4, and IGHGP), IgE (IGHEP1 and IGHE), and IgA (IGHA1 and IGH2).

Interestingly, the IGHG genes, to our knowledge, are being comprehensively characterized for the first time. Previous studies have relied on a single assembly, Kamilah_GGO_v0, which included only four IGHC genes: IGHM, IGHD, IGHG3A, and IGHG1 (45, 46).

The gorilla IGH locus includes three IGHG3 genes, which differ from each other by their number of hinge exons (two or five for IGHG3A depending on the allelic polymorphism, two for IGHG3B, and four for IGHG3C). The gorilla C-CLUSTER shows a highly similar organization to the human one. The main difference is the addition of two IGHG3 genes, presumably IGHG3A and IGHG3B since the human IGHG3 and the gorilla IGHG3C share the same number of hinge exons and almost 100% of identity in all their exon sequences.

3.2 *G. gorilla gorilla* IGK locus

3.2.1 Localization and description of IGK locus

The gorilla IGK locus has a reverse orientation (REV) on chromosome 2A for the three assemblies Kamilah_GGO_v0, NHGRI_mGorGor1-v1.1-0.2.freeze_mat, and NHGRI_mGorGor1-v1.1-0.2.freeze_pat, whereas the Susie3 IGK locus is forward (FWD) (Table 1). However, the REV or FWD orientation does not appear to

affect the genomic structure, gene organization, or gene functionality. According to IMGT rules for quality assessment of IG and TR loci in genome assemblies, the IGK locus was satisfactory for genomic annotation across all four assemblies.

Two IMGT flanking genes “IMGT bornes” (30) were identified for the IGK locus delimitation of many species (see the section “Locus bornes: IGK locus 5’ and 3’ bornes”¹⁴ of IMGT Repertoire), including the gorilla. The 5’ end IMGT flanking gene, “IMGT borne”, PAX8 (NCBI Gene ID: 101137174) is 125 kb upstream of IGKV1-49, the most 5’ gene in the locus. The 3’ end IMGT flanking gene, “IMGT borne”, RPIA (NCBI Gene ID: 101148377) is 200 kb downstream of IGKC, the most 3’ gene in the locus. “IMGT bornes” were identified in the four annotated assemblies.

The IGK locus extends from 10 kb upstream, IGKV1-49, to 10 kb downstream, IGKC. The differences in IGK locus size observed for the four assemblies (varying from 698 kb to 1,374 kb) were observed between the genes IGKV2-29 and IGKV3-31. The region between these two genes seems to not comprise IGK genes or any other genes. The IGK locus comprises a cluster of 41 to 44 V genes (V-CLUSTER) for a large part and a cluster of 5 J and 1 C genes (J-C-CLUSTER) (Table 1; Supplementary Figures 5-8; Supplementary Table 2).

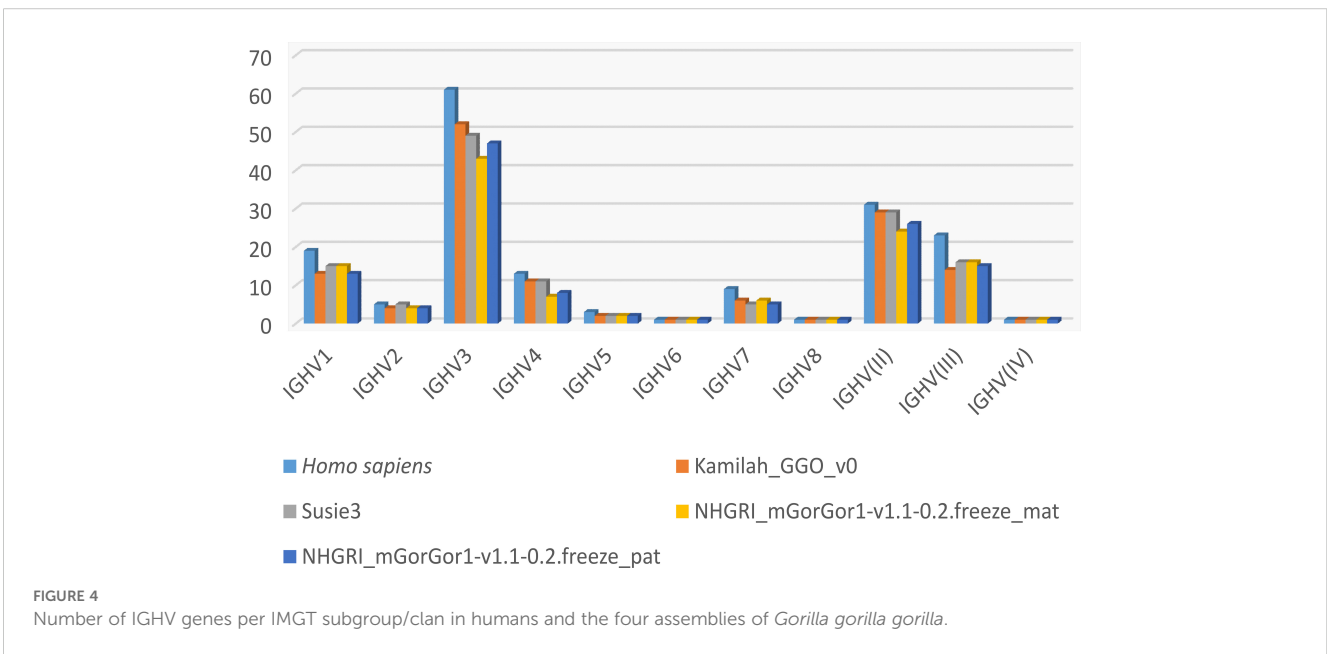
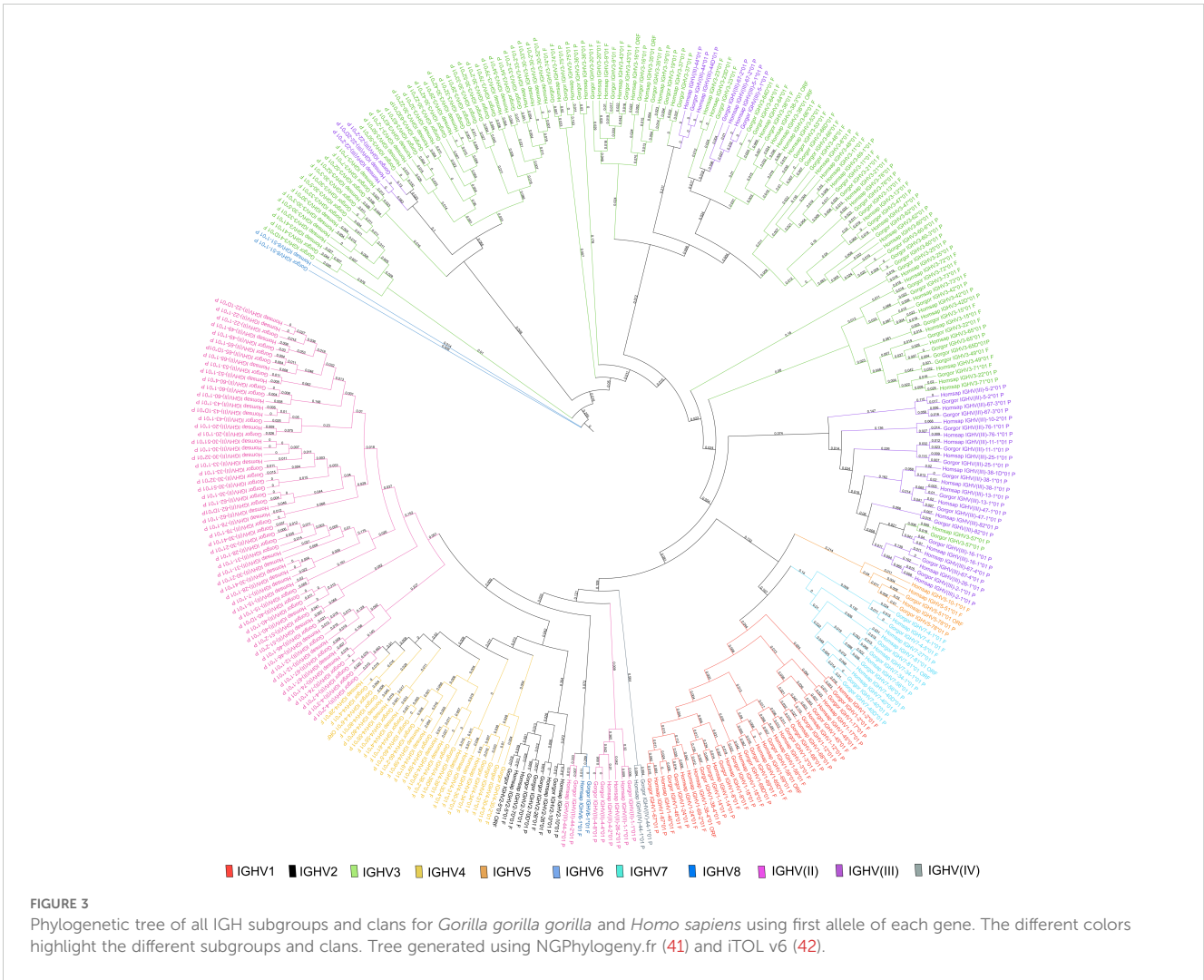
3.2.2 IG gene organization in the gorilla IGK locus

3.2.2.1 IGKV gene cluster

A total of 44 IGKV genes were identified from the annotation of the four assemblies; 31 of them show polymorphic alleles, and in total, 92 IGKV alleles were characterized (Figure 2; Supplementary Table 2).

¹³ <https://www.imgt.org/IMGTrepertoire/LocusGenes/genetable/autotable.php>

¹⁴ <https://www.imgt.org/IMGTrepertoire/LocusGenes/bornes/bornesIGK.html>



The IGKV genes were classified into seven subgroups defined according to the IMGT-ONTOLOGY (36) and their sequence similarity with the IGKV human subgroups. The phylogenetic tree in **Supplementary Figure 13**, constructed from the first allele of all gorilla and human IGKV genes, displays the distances between the IGKV genes of both species. It shows that the gorilla genes are grouped in the same clade as their human counterpart genes.

The number of IGKV genes in gorillas is slightly over half that of the human IGKV gene number, with 76 localized IGKV genes in the main human locus. The human IGK locus comprises a proximal V-CLUSTER (p) of 40 IGKV genes, and a distal V-CLUSTER (d) of 36 genes, from 3' to 5' (1, 47). The first eight gorilla IGKV genes starting from the 3' to 5' of the IGK locus have an extremely close organization, gene order, sequence similarity, and gene orientation (IGKV4-1 and IGKV5-2 have opposite orientation in the locus, as well as in humans) to that of the human proximal (p) IGK V-CLUSTER. The other gorilla IGKV genes are mostly closer to the human distal (d) IGK V-CLUSTER (**Supplementary Table 2**). Moreover, counterparts of the human IGKV6D-41, IGKV1D-42, and IGKV1D-43 were identified in gorilla assemblies, whereas they were not in the human proximal (p) IGK V-CLUSTER. Conversely, it should also be noted that there is no counterpart of IGKV1-9 (located in the human proximal IGK V-CLUSTER) in the gorilla assemblies, nor is there a corresponding duplicated gene in the human distal IGK V-CLUSTER.

Therefore, the IGK gene nomenclature of the human proximal IGK V-CLUSTER was assigned to all gorilla IGKV genes, including IGKV6-41, IGKV1-42, and IGKV1-43 genes, according to their orthology, and especially since gorilla does not have duplicated IGK V-CLUSTER. For additional genes, names were assigned according to IMGT nomenclature rules.

The genes between IGKV1-44 and IGKV1-49 are additional in the gorilla IGK locus compared to the human one, and the positional nomenclature was adopted by incrementation of the position number on the locus.

Almost 1/3 of gorilla IGKV genes (16, 17) belong to the IGKV1 subgroup, which includes the highest number of functional genes, and the other 1/3 (13–15) belong to the IGKV2 subgroup, which includes the highest number of pseudogenes (**Supplementary Table 6; Supplementary Figure 15**).

3.2.2.2 IGKJ and IGKC gene clusters

The annotation of the four gorilla IGK assemblies allowed us to highlight five IGKJ genes belonging to five sets—1, 2, 3, 4, and 5—one gene for each set, and one unique constant IGKC gene (**Supplementary Table 2**). In examining the assemblies, these genes appear to be minimally or not at all polymorphic: only one additional allele was shown for the IGKJ gene and none for IGKC (**Figure 2**).

3.3 G. gorilla gorilla IGL locus

3.3.1 Localization and description of IGL locus

The IGL locus for the four selected gorilla assemblies is delimited by flanking genes, “IMGT bornes” (30) identified in other species (see the section “Locus bornes: IGL locus 5' and 3'

bornes”¹⁵ of IMGT Repertoire) (**Table 1**). The 5' end IMGT IGL Locus borne is the TOP3B gene (NCBI Gene ID: 101141903), which is in reverse orientation. The 3' end IMGT IGL Locus borne is the RSPH14 gene (NCBI Gene ID: 101130781). The IGL locus extends from 10 kb upstream of the most 5' gene in the locus, IGLV(I)-70-1, to 10 kb downstream of the most 3' gene in the locus, IGLC7, and comprises a cluster of 76 to 86 V genes (V-CLUSTER) and a cluster of 7 to 8 IGLJ and C genes (J-C-CLUSTER). Interestingly, six known RPI genes within the human IGL locus were also identified in gorilla assemblies (**Table 1; Supplementary Figures 9-12; Supplementary Table 3**).

3.3.2 IG gene organization in the gorilla IGL locus

3.3.2.1 IGLV gene cluster

The analysis of the four assemblies allowed the identification of 86 IGLV genes in total. From gene annotations, 53 genes show allelic polymorphism, and in total, 163 IGLV alleles were characterized (**Figure 2; Supplementary Table 3**).

The IGLV genes were classified into 11 subgroups and seven clans defined according to IMGT-ONTOLOGY (36) and to their sequence similarity with the human IGLV subgroups. The phylogenetic tree of **Supplementary Figure 14**, built from the first allele of all gorilla and human IGLV genes, displays the distance between the IGLV genes of both species and shows that the gorilla genes are grouped in the same clade with their human counterpart gene. The clans IGLV(I), IGLV(II), and IGLV(V) are interspersed between subgroups because of the sequence similarity and the gene nomenclature according to the “IMGT IGL clan tree”¹⁶.

As for humans and other non-human primates, the gorilla IGLV3 subgroup gathers the highest number of genes (20 to 24 depending on the assembly), with approximately the same number of functional genes and pseudogenes (**Supplementary Table 7; Supplementary Figure 16**).

The comparison of the IGL locus from the four assemblies with the IGL locus organization of humans highlights common features between the four assemblies. Indeed, two blocs of human IGLV genes are absent in the gorilla locus (highlighted in yellow in **Supplementary Table 3**): a bloc of seven genes [namely, IGLV(VII)-41-1, IGLV1-41, IGLV1-40, IGLV5-39, IGLV(I)-38, IGLV5-37, and IGLV1-36] and a bloc of five genes [namely, IGLV(IV)-66-1, IGLV(V)-66, IGLV(IV)-65, IGLV(IV)-64, and IGLV(I)-63]. The availability of assemblies from more individuals should help to confirm if this observation corresponds to a gorilla-specific feature. Another gorilla-specific attribute would correspond to the 16 IGLV genes (highlighted in green in **Supplementary Table 3**) present in the four gorilla assemblies but not in humans. We also identify a potential CNV between IGLV3-24-2 and IGLV3-27, an area of eight IGLV genes not identified in all four assemblies.

We noticed that six genes—IGLV(I)-34-1, IGLV2-34, IGLV2-33, IGLV3-32, IGLV3-31, and IGLV3-30—were not identified in

¹⁵ <https://www.imgt.org/IMGTrepertoire/LocusGenes/bornes/bornesIGL.html>

¹⁶ <https://www.imgt.org/IMGTIndex/IGLVclans.php>

the Kamilah_GGO_v0 assembly. This seems to be linked to the presence of a gap of 47 kb in this position, and this cannot be considered a potential CNV.

3.3.2.2 IGLJ-C gene cluster

The gorilla IGL J-C-CLUSTER is composed of seven tandems of IGLJ and IGLC genes or eight (IGLJ2A and IGLC2A for NHGRI_mGorGor1-v1.1-0.2.freeze_pat only, which is considered potential CNVp2 in gorilla). The IGLJ genes show a very low allelic polymorphism (only two alleles for IGLJ5).

4 Discussion

The identification of the gorilla IG genes and alleles, along with the characterization of their genomic organization detailed in the present study, increases our knowledge of the genetics of the adaptive immune response in jawed vertebrates. Additionally, it provides interesting clues regarding the molecular evolution and conservation of gorilla IG loci among primates, as well as the individual variations within the population.

To detect potential evolutionary events in germline DNA sequences of gorilla IG loci, we relied on a gorilla-human comparative genomics study. Whatever the loci of the three individuals and the four NCBI assemblies (Table 1), the gorilla IG loci have retained a structure close to the related locus in humans with approximately the same number of genes (except for IGK if we count the number of genes in the proximal and distal copies). Sequences of both species present high similarity, which is closely correlated with the taxonomic relationship. The speciation event led to the conservation of orthologous genes in gorillas, and according to IMGT gene nomenclature, IG gorilla genes were assigned the names of their orthologous human IG genes, if any. In addition, orthologous gene positions in the loci and the use of the IMGT positional nomenclature for gorilla genes were also confirmed with the detection of flanking genes, “IMGT bornes” (30), when they exist (IGL and IGK loci), and with highly conserved RPI in the IGH and IGL loci. These RPI sequences are conserved in all mammals and used as markers in the locus (30).

Comparison of the gorilla IG loci from the four assemblies (Kamilah_GGO_v0, Susie3, NHGRI_mGorGor1-v1.1-0.2.freeze_mat, and NHGRI_mGorGor1-v1.1-0.2.freeze_pat) highlights genomic variations that have been observed exclusively in gorilla species, which may suggest that the genome is accumulating unique variations depending on each individual. The IG genomic sequences of the four assemblies were selected according to IMGT rules for assessment of IG and TR loci in genome assemblies. It is worth mentioning that the NHGRI_mGorGor1-v1.1-0.2.freeze_pri, the NCBI “representative genome” of Western lowland gorilla since March 2023, includes exactly all IG genes and alleles of NHGRI_mGorGor1-v1.1-0.2.freeze_mat (from which the IG loci were analyzed), and therefore, the IG loci of this assembly are not detailed in the current article.

After analyzing all previous assemblies, two additional ones were published online—Kamilah_GGO_hifiasm-v0.15.2.pri and Kamilah_GGO_hifiasm-v0.15.2.alt—which were acquired from

the same individual, Kamilah, utilizing PacBio Sequel technology and HiFiasm v. 0.15.2 assembly method. The latter two assemblies are at the contig level; therefore, they were not fully analyzed and included in this article because they are not on the chromosome level.

4.1 V-GENE multigene families, allelic polymorphism, gene insertion/deletion, and CNV identification in Western lowland gorilla IG loci

The diversity of the IG variable domains is partly generated by the repertoire of large numbers of variable (V) genes in the germline DNA (8). This is especially true for the V genes of the heavy chain, which are more numerous than those of light chains in most species, which is the case in Western lowland gorilla. As mentioned in Reference (8), the reason behind the expansion and contraction of the IGHV multigene families in jawed vertebrates is still poorly understood. The duplication and divergence events in IG loci are governed by different patterns driven by natural selection.

Comparison of V genes by multiple alignments between the annotated loci of the different assemblies revealed the duplication of certain genes, as well as the divergence of duplicated genes that occur during locus evolution. In both cases, the genes are considered phylogenetically related. The more duplicated genes that occur, the more nonfunctional genes are produced in the IGHV multigene families (6). Importantly, these pseudogenes are carefully considered in the IMGT annotation of germline DNA, as they provide precious clues to the organism’s evolution.

4.1.1 IMGT subgroups of IGH, IGK, and IGL variable genes

IMGT subgroup names of non-human primates have been assigned by homology with those of humans. The classification of gorilla variable genes into IMGT subgroups highlights the abundance of certain subgroups: IGHV3 for IGH, IGKV1 for IGK, and IGLV3 for IGL (Figure 4; Supplementary Tables 1-3; Supplementary Figures 15, 16). Interestingly, this result is also observed for other non-human primate species, such as rhesus monkey (*Macaca mulatta*) (18), Sumatran orangutan (*Pongo abelii*), Bornean orangutan (*Pongo pygmaeus*), and even IGH and IGL of ring-tailed lemur (*Lemur catta*), a more distant primate species in the taxonomy classification (data available at “Locus gene repertoire per IMGT annotated assembly”¹⁷).

4.1.2 Allelic polymorphism

The gorilla IG genes are shown to be polymorphic (Supplementary Table 1-3); as for other primate species, the different alleles result from nucleotide substitutions and/or nucleotide insertions or deletions, which may lead to a modification of the gene functionality. For some alleles of genes,

¹⁷ https://www.imgt.org/IMGTrepertoire/LocusGenes/locusdesc/assembly_compare.php

e.g., IGHV7-40, IGHV7-40D, IGKV2-23, and IGKV2-38, the functionality was altered due to the insertion of repeated foreign IG DNA sequences, known as “Repeat regions”, mostly LINE and SINE families. The proportion of these regions is over 60% in mammals (48).

The characterization of the allelic polymorphisms was based on the IMGT unique numbering (34, 35) for an easy comparison of codons and amino acid sequences of the V, D, J, and C regions or exons (34). The dynamic gene tables per IMGT group and species¹⁸ list the alleles of IG genes and their corresponding functionality. In the gene table, a scoring system based on one to three stars indicates that a given allele was identified in one, two, or more genomic sequences. In the context of the evolution of high-throughput sequencing technologies, more than one star would confirm the existence of the alleles and eliminate suspicion of sequencing errors. Following the analysis of the four assemblies, 63% of IGH, 57% of IGL, and 64% of IGK genes are shown to be polymorphic. The higher the number of annotated assemblies in the future, the more accurate this estimation will be.

4.1.3 Gene insertion/deletion and CNVs in the IGHV cluster

The analysis of the IGH locus within the four assemblies of gorillas allowed us to report variations in the gorilla genome between individuals, in particular CNVs (Figure 5; Supplementary Table 1). Among these, some CNVs have already been described in human and some non-human primates (49), indicating that these CNVs may not be specific to the human species. Indeed, Gazave and colleagues (50) observed that the majority of CNVs are not species specific, and they are consistent with species phylogenetic relationships. Shared CNVs may be the result of ancient structural polymorphism retention, as well as high segmental duplication activity, which facilitates recurrent loss or creation of new copies via non-allelic homologous recombination (50).

According to the human and gorilla genes organization in the IGH locus, this could be illustrated by the human CNV3 (30) (represented in the “Human (*Homo sapiens*) IGH CNV3 IMGT”¹⁹ web page). The counterpart of this CNV was also identified in gorillas with the CNV3 form (gene content of the CNV between the 3' and 5' limit) C in Kamilah_GGO_v0 -haplotype 1- and two new CNV3 forms, not described in humans, called forms H and I, apparently gorilla specific.

It is worth noting that additional human CNVs, namely, CNV1, CNV6, and CNV7, could be identified in gorilla IGH loci (with variation in the number of genes), with gorilla-specific CNV forms (Supplementary Table 1).

New sets of IG genes have been identified in gorilla loci compared to humans: five of them could be associated with new gorilla-specific CNVs, called potential CNVp1 to CNVp5 (Supplementary Table 1). However, another set composed of the

six IGHD genes absent from Kamilah_GGO_v0 assembly because of the 20-kb gap is not proposed as a potential CNV.

4.2 IG constant genes and CNVs

4.2.1 IGH C-GENE class and subclass characterization

The five IG classes (namely, IgA, IgE, IgD, IgM, and IgG) are characterized by different heavy chain constant regions, coded by the constant genes of the IGH locus (51). In this study, we highlighted a highly similar IGHC gene organization between gorillas and humans (Supplementary Table 2; Supplementary Figures 1-4). This corroborates previous work describing gorillas, chimpanzees, and human and orthologous genes (45). However, even if the presence of several IGHC was already mentioned (45, 46), to our knowledge, this is the first study that identifies and characterizes the nine distinct IGHC genes and in particular the six IGHC genes. Indeed, the three IGHC3 genes, IGHC3A, IGHC3B, and IGHC3C, are present in three of four assemblies. This would confirm that duplications continued to occur especially in the clade of IgG3 where gorillas and chimpanzees created an additional IgG, which reflects evolutionary instability in the locus. The gorilla IgG3 isotype is characterized by three constant domains and a variant number of hinge regions, from two to five (two to five for IgG3A, two for IgG3B, and four for IgG3C) (see Dynamic gene tables per IMGT group and species). The hinge in the IgG3 immunoglobulin class is therefore longer than the hinge regions of the other IgG subclasses (51), except for IGHP. According to the number of hinge regions (and to high similarity of domain exons), the gorilla IGHC3C gene seems to be the most closely matched to the human IGHC3.

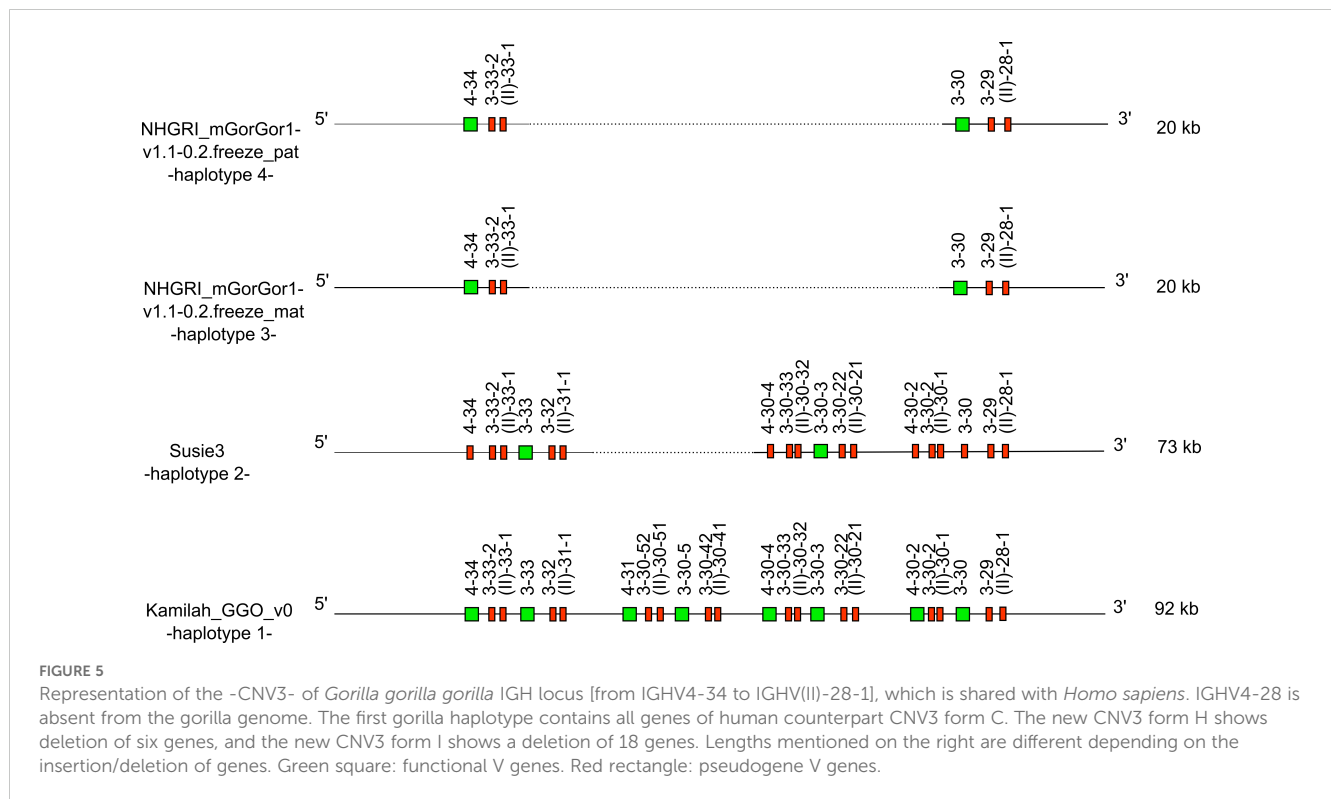
Sequences of IGHC2, IGHC4, IGHE, and IGHA2 were not detected within the IGH loci of the Kamilah_GGO_v0 nor those of Susie3. In the latest assembly, we found the four genes on contig CYUI03001141.1²⁰ (data not shown) associated with the related bioproject but not assembled on chromosome 14. These four genes were found and annotated within the IGH locus in both NHGRI_mGorGor1-v1.1-0.2.freeze_mat and NHGRI_mGorGor1-v1.1-0.2.freeze_pat haploid genomes.

Humans, chimpanzees, and gorillas seem to share a common ancestral duplication of the IGHC, IGHE, and IGHA genes (52), which likely had taken place in their common ancestor. Therefore, the IGHE and IGHEP1 genes were linked to IGHA2 and IGHA1, respectively, in the gorilla genome (52). In order to confirm the correct gene name assignment of the IGHEP1, IGHA1, IGHE, and IGHA2 genes, the characteristic length of IGHA genes was taken into account. On the one hand, our results contradict those reported in Reference (52) regarding the hinge region length of the gorilla IGHA1 compared to those of humans and chimpanzees: we noticed the presence of duplication in the hinge region of gorilla IGHA1 that also occurred in humans and chimpanzees. We confirm that the hinge of the third allele of the gorilla is shorter because of the deletion of two nucleotides, leading to a frameshift in the reading

18 <https://www.imgt.org/IMGTrepertoire/LocusGenes/genetable/autotable.php>

19 https://www.imgt.org/IMGTrepertoire/LocusGenes/locus/human/IGH/CNV/Hu_IGHCNV3.html

20 <https://www.ncbi.nlm.nih.gov/nuccore/CYUI03001141.1>



frame (Figure 6A). On the other hand, we concur with the assumption made by the same scientific team that the IGHA2 gene was derived from the prototype IGHA1 by the 15-bp deletion in the hinge region, before its duplication, which seems to have occurred before the divergence of the three species (humans, chimpanzees, and gorillas) (Figure 6B).

Two IGHE genes were annotated on the main IGH locus of gorilla—IGHE and IGHEP1—which is truncated in 5'. It seems that among the hominoids, only the gorilla and human genomes contained three IGHE genes (54): two in the main locus and one outside. Note that the detected IGHE gene outside the main locus of the gorilla (data not shown) is the human IGHEP2 counterpart (processed gene outside the main IGH locus in the human genome).

It should be noted that the IGHC genes from *G. gorilla* species have been previously annotated and published on the IMGT site. However, because of unknown subspecies and partial and/or non-identical sequences, they were not reassigned according to this study (the *G. gorilla* IGHG3 could not be assigned to subtype IGHG3A, IGHG3B, or IGHG3C, and new allele numbers were assigned to the genes from the four assemblies if the genes were already published in the “Gene table: Western gorilla (*G. gorilla*) IGHC”²¹).

4.2.2 IGLJ and IGLC genes

An additional J-C-CLUSTER was identified in the gorilla IGL locus of the NHGRI_mGorGor1-v1.1-0.2.freeze_pat assembly compared to available IMGT annotated human assemblies. The

study (55) comparing IGL sequences between different human populations revealed that some human populations could have up to four additional IGLC genes, most likely linked to a junction gene, localized between the IGLC2 and IGLC3 genes (see “Locus representation: human (*H. sapiens*) IGL”²² on IMGT web site). As found in the same location in the gorilla counterparts, this represents a form of CNV, with 99% and 100% identity between IGLC2 and IGLC2A, and between IGLJ2 and IGLJ2A, respectively.

4.3 Gorilla and human IGK locus analyses

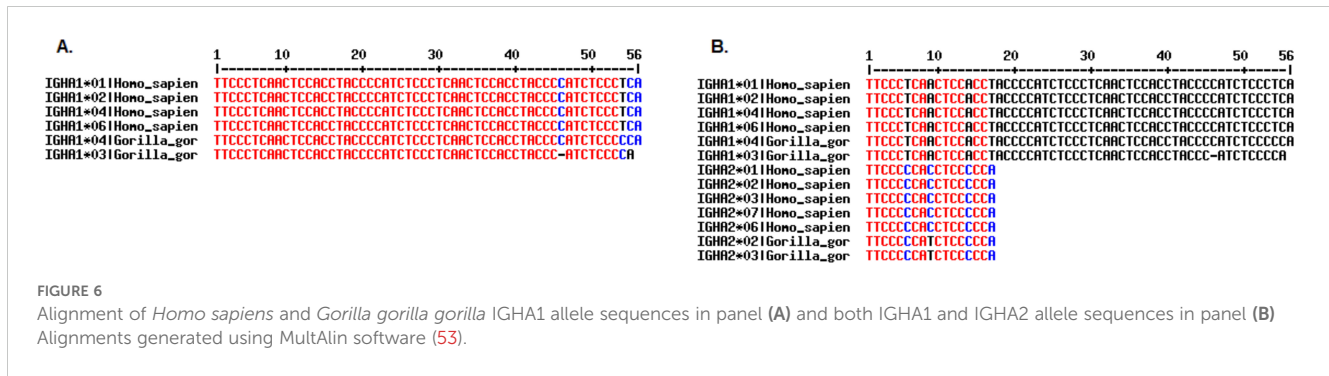
The IGK locus has a reverse orientation on chromosome 2A in three assemblies but is forward on chromosome 2A of the Susie3 assembly. We noticed that this unexpected locus orientation was also observed for the dog, *Canis lupus familiaris*: the IGK locus is REV for the CanFam3.1 and FWD for the Basenji_breed-1.1, both assemblies annotated in IMGT (see “Locus in genome assembly: dog (*C. lupus familiaris*) IGK locus”²³).

Indeed, dog and gorilla assemblies were built using the comparison with the human one (the human IGK locus is composed of a proximal IGK in REV orientation and a duplicated part in FWD orientation on chromosome 2). As

²² <https://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=locus&species=human&group=IGL>

²¹ <https://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=Gorilla&group=IGHC>

²³ <https://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=locusAssembly&species=dog&group=IGK>



neither the dog nor the gorilla shows a duplicated part in their IGK locus, this individual change in the IGK locus orientation could be linked to a methodology artifact.

The gorilla IGK locus contains six additional IGKV genes in the 5' side of the V-CLUSTER with no identified human counterparts up to now.

Our results, which indicate the detection of IGK genes only between gorilla IGK 5' and 3' "IMGT bornes", confirm the existence of one unique IGK locus. As cited in Reference (56), no indication of duplication within the IGK locus was obtained in establishing the *Pan troglodytes* and the *G. gorilla* maps. In contrast, the human IGK locus has two V-CLUSTERS in inverse orientation to each other, which are very similar but not identical, called the proximal (p) and distal (d) loci (57).

Our findings show that the genes of gorilla IGK locus present a high percentage of identity with human genes of the distal IGK V-CLUSTER and similar structural organization especially since we found in gorilla three genes corresponding to the three additional human gene counterparts of the distal V-CLUSTER, which have no duplicate equivalent on the human proximal V-CLUSTER: IGKV6-41, IGKV1-42, and IGKV1-43 (Supplementary Table 2). The divergence between proximal and distal V-CLUSTER is largely due to points of mutations (58) involving deletions in some regions on the proximal locus which must have occurred after duplication of the locus (59, 60), indicating that this duplication is an evolutionary process (58). The absence of two V-CLUSTER in the IGK locus in chimpanzees and gorillas means that the duplication in the human IGK locus occurred after the branch-point human and great ape evolution (56).

4.4 *G. gorilla gorilla* chromosome nomenclature

Up to January 8, 2024, non-human primates' chromosomes were named by homology with the ones of humans. The common ancestor of gorillas, chimpanzees, and humans had 24 pairs of chromosomes (61). Great apes have conserved the same number of diploid chromosomes, whereas modern humans possess 23 pairs ($2n = 46$) due to a telomeric fusion of chromosomes 2A and 2B. Most chromosomes appear to be similar between the three species, with the remaining differences between chromosomes consisting of

inversions of chromosome segments and variations in constitutive heterochromatin (61). Based on this nomenclature, we were able to localize the three IG loci on the same chromosomes as humans: the IGH locus on chromosome 14, the IGK locus on chromosome 2A in the gorilla, and the locus IGL on chromosome 22.

Since January 8, 2024, the non-human primate chromosome pairs were updated and renamed from 1 to 24. For gorilla, only assemblies of KB3781 individual have been updated (NHGRI_mGorGor1-v2.0_pri, NHGRI_mGorGor1-v2.0_mat, and NHGRI_mGorGor1-v2.0_pat). Therefore, the IGH locus now resides on chromosome 15, the IGK locus on chromosome 12, and the IGL locus on chromosome 23. In the context of phylogenetic studies between apes and humans, and gorilla individuals, the previous version of assemblies was more appropriate, in our opinion, in terms of the close phylogenetic relationship between gorillas and humans and their common ancestor.

We strongly believe that this sort of important change should be taken after consultation with the scientific community and clear prior communication before implementation.

4.5 Assemblies of "Kamilah" individual

The Kamilah_GGO_v0, Kamilah_GGO_hifiasm-v0.15.2.pri, and Kamilah_GGO_hifiasm-v0.15.2.alt assemblies (Figure 1) originate from the same individual, Kamilah. The analysis of Kamilah_GGO_v0 (oldest Kamilah assembly) IG loci is an integral part of this study, while the analysis of Kamilah_GGO_hifiasm-v0.15.2.pri and Kamilah_GGO_hifiasm-v0.15.2.alt in the NCBI assemblies at the "Contig" level were not integrated because they do not correspond to our assembly selection criteria for this study.

Upon a preliminary analysis of the IG loci in the two most recent assemblies, similar gene organization within the same locus was confirmed as expected. However, we observed variations in gene number within the IGH and IGL loci but not in the IGK locus (Figure 7).

Considering the origin of the data, the source of differences could be linked to the biological material and/or to the sequencing technologies:

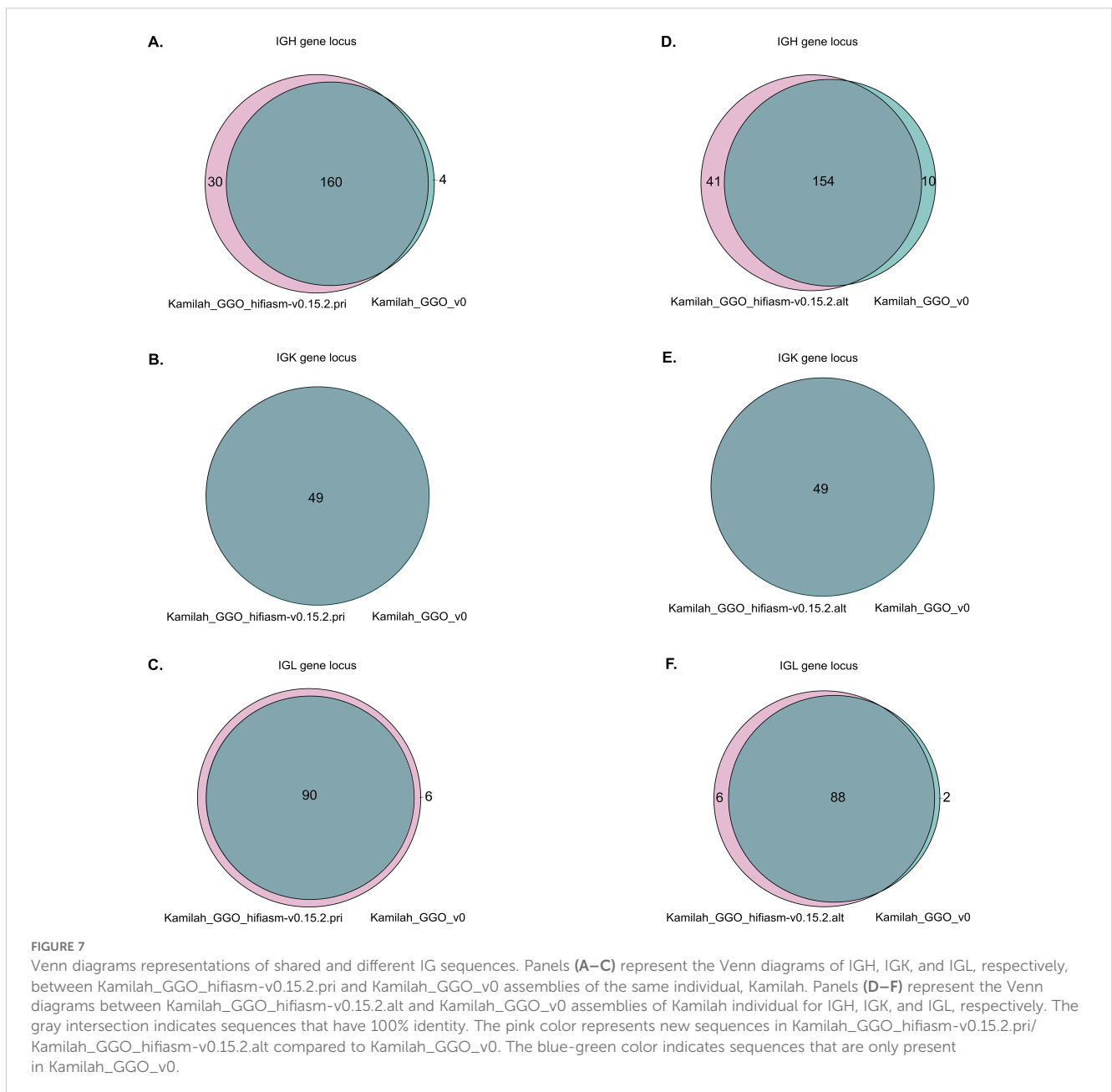
-Given that the Kamilah_GGO_v0 genome was obtained from a primary cultured fibroblast cell line and that both Kamilah_GGO_

hifiasm-v0.15.2.pri and Kamilah_GGO_hifiasm-v0.15.2.alt were obtained from the cell line, the different tissue types influence genomic stability. On the one hand, cell lines are frequently derived from a single cell, yet they may accrue genetic changes over time due to extended cultivation. Primary cells, on the other hand, may better reflect the individual's genetic composition, although they may comprise a variety of cell types and are subject to culture changes.

-Depending on the sequencing and assembly methods, the identified sequence variations may be linked to methodological parameters such as sequencing coverage, read depth, mapping quality, assembly contiguity, and accuracy. Illumina technology produces shorter, lower-quality reads than PacBio technology, which produces longer reads (62). PacBio Sequel technology offers a higher consensus accuracy than PacBio RS II (63).

The Kamilah_GGO_hifiasm-v0.15.2.pri genome has more genes than the Kamilah_GGO_v0 genome. This is because Kamilah_GGO_v0 is missing 15 IGHD genes (due to a gap at this position) and nine IGHC genes in the IGH locus. Additionally, the Kamilah_GGO_hifiasm-v0.15.2.pri genome contains additional IGH and IGL V genes, which are not present in Kamilah_GGO_v0.

However, several duplications (in V-CLUSTER and C-CLUSTER in IGH locus) seem to occur in the Kamilah_GGO_hifiasm-v0.15.2.alt assembly. These would be more likely the result of sequencing and/or assembly errors, although the Hifiasm assembly technique (utilized for the two recent assemblies of Kamilah) has a clear advantage over the other assemblers investigated in Reference (64), including the FALCON assembly method, the one used for Kamilah_GGO_v0.



These preliminary studies allowed us “to fill the gaps” in the D-CLUSTER of the IGH locus of Kamilah_GGO_v0 and to confirm the existence of nine additional IGHC genes in the Kamilah individual. However, an in-depth and complete analysis would be needed to interpret the meaning of differences between the three concerned assemblies.

5 Conclusion

By deciphering the immunoglobulin genes at the three IG loci (IGH, IGK, and IGL) from four Western lowland gorilla (*G. gorilla gorilla*) NCBI genome assemblies, IMGT[®] provides a consistent overview of the organization and description of these loci and the potential individual variations in this closely related primate to humans.

Due to the highly similar organization of gorilla and human loci and the high percentage of identity between IG genes in gorillas and humans, the IMGT names of gorilla IG genes were mostly assigned according to their human counterparts.

The IG loci and the gene characterization, thanks to IMGT gene nomenclature and IMGT standards, highlighted characteristics of the gorilla genome:

As in the human IGH locus, the gorilla IGH locus shows the greatest variability between individuals in terms of gene content. Several known human CNVs were identified in the gorilla IGH locus, along with new forms, as well as other potentially new CNVs called CNVp until their confirmation in other assemblies.

The analysis of the organization of IG constant genes in the IGH locus from several individuals helped to better estimate the number of IGHC genes, which had been previously underestimated based on a single assembly (46), particularly with the characterization of three IGHC3 genes.

The IGK locus is remarkably homogeneous in the four assemblies: it is characterized by the absence of IGKV locus duplication, which occurred in the human IGK locus; in addition, the IGKV gene cluster seems to be closer to the distal human V locus.

The IGL locus comprises a CNV in the J-C-CLUSTER, which was suspected in the human IGL locus but had not been shown in the IMGT-annotated IGL locus until now.

The analysis of these loci generated a large amount of expertly curated data from the three gorilla individuals, which are distributed through the IMGT website resources, databases, tools, and web resources compiled in **Supplementary Table 4**. Although data from three individuals cannot reflect those of an entire population, they have enriched our immunogenetics knowledge of this species, a closely related primate to humans, and will continue to evolve with the publication and expertise of new genome assemblies based on improved sequencing technologies and data from an increasing number of individuals.

The analysis of immunogenetics data is crucial in current immunology research. Studying great apes like gorillas, which are central to the Hominoidea group, offers valuable insights into primate evolution.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: NCBI (<https://www.ncbi.nlm.nih.gov/>): GCA_008122165.1, GCA_900006655.3, GCA_028885495.1, GCA_028885475.1.

Author contributions

CD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. GF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review & editing. JJ-M: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review & editing. VG: Conceptualization, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SK: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. CD was funded by a doctoral contract from the Algerian Excellence Scholarship, Algerian Ministry of Higher Education and Scientific Research. We acknowledge the support of Immun4Cure IHU “Institute for innovative immunotherapies in autoimmune diseases” (France 2030/ANR-23-IHUA-0009). IMGT[®] is substantially supported by the Centre National de la Recherche Scientifique (CNRS) and the University of Montpellier.

Acknowledgments

We thank all members of the IMGT[®] team for their expertise and constant motivation. We recognize the groundbreaking contributions of Marie-Paule Lefranc, the founder of IMGT[®]. IMGT[®] is member of the French Infrastructure Institut Français de Bioinformatique (IFB) as well as member of BioCampus, MAbImprove, and IBiSA. This work was granted access to the High Performance Computing (HPC) resources of Meso@LR and of Centre Informatique National de l'Enseignement Supérieur (CINES), to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), and to Institut du développement et des ressources en informatique scientifique (IDRIS) under the allocation 036029 (2010–2024) made by GENCI (Grand Équipement National de Calcul Intensif).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2024.1475003/full#supplementary-material>

References

- Lefranc MP, Lefranc G. *The Immunoglobulin FactsBook*. London, UK: Academic Press (2001). 472 p.
- Lefranc MP, Lefranc G. *The T Cell Receptor FactsBook*. London, UK: Elsevier (2001). 413 p.
- Marshall JS, Warrington R, Watson W, Kim HL. An introduction to immunology and immunopathology. *Allergy Asthma Clin Immunol*. (2018) 14:49. doi: 10.1186/s13223-018-0278-1
- Tonegawa S. Somatic generation of antibody diversity. *Nature*. (1983) 302:575–81. doi: 10.1038/302575a0
- Giudicelli V, Lefranc MP. Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics*. (1999) 15:1047–54. doi: 10.1093/bioinformatics/15.12.1047
- Das S, Nozawa M, Klein J, Nei M. Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates. *Immunogenetics*. (2008) 60:47–55. doi: 10.1007/s00251-007-0270-2
- Hirano M, Das S, Guo P, Cooper MD. The Evolution of Adaptive Immunity in Vertebrates. In: *Advances in Immunology*. Oxford, UK: Elsevier's Science & Technology Rights Department (2011). p. 125–57. Available online at: <https://linkinghub.elsevier.com/retrieve/pii/B9780123876645000042>.
- Tanaka T, Nei M. Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Mol Biol Evol*. (1989) 6:447–59. doi: 10.1093/oxfordjournals.molbev.a040569
- Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res*. (2015) 43:D413–22. doi: 10.1093/nar/gku1056
- Manso T, Folch G, Giudicelli V, Jabado-Michaloud J, Kushwaha A, Nguefack Ngoune V, et al. IMGT® databases, related tools and web resources through three main axes of research and development. *Nucleic Acids Res*. (2022) 50:D1262–72. doi: 10.1093/nar/gkab1136
- Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT(R), the international ImMunoGeneTics information system(R). *Nucleic Acids Res*. (2009) 37:D1006–12. doi: 10.1093/nar/gkn838
- Lefranc MP. Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. *Front Immunol*. (2014) 5:22/abstract. doi: 10.3389/fimmu.2014.00022/abstract
- Grow DA, McCarrey JR, Navara CS. Advantages of nonhuman primates as preclinical models for evaluating stem cell-based therapies for Parkinson's disease. *Stem Cell Res*. (2016) 17:352–66. doi: 10.1016/j.scr.2016.08.013
- Bailey JA, Eichler EE. Erratum: Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*. (2006) 7:898–898. doi: 10.1038/nrg1997
- Symmons O, Varadi A, Aranyi T. How segmental duplications shape our genome: recent evolution of ABCC6 and PKD1 mendelian disease genes. *Mol Biol Evol*. (2008) 25:2601–13. doi: 10.1093/molbev/msn202
- Scally A, Duthell JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*. (2012) 483:169–75. doi: 10.1038/nature10842
- Rogers J, Gibbs RA. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat Rev Genet*. (2014) 15:347–59. doi: 10.1038/nrg3707
- Nguefack Ngoune V, Bertignac M, Georga M, Papadaki A, Albani A, Folch G, et al. IMGT® Biocuration and analysis of the rhesus monkey IG loci. *Vaccines*. (2022) 10:394. doi: 10.3390/vaccines10030394
- Wilming LG, Hart EA, Coggill PC, Horton R, Gilbert JGR, Clee C, et al. Sequencing and comparative analysis of the gorilla MHC genomic sequence. *Database*. (2013) 2013:bat011. doi: 10.1093/database/bat011/330685
- Van Loghem E, De Lange G. Immunoglobulin epitopes in primates. *Vox Sanguinis*. (1979) 37:329–37. doi: 10.1111/j.1423-0410.1979.tb02312.x
- Pégorier P, Bertignac M, Chentli I, Nguefack Ngoune V, Folch G, Jabado-Michaloud J, et al. IMGT® Biocuration and comparative study of the T cell receptor beta locus of veterinary species based on homo sapiens TRB. *Front Immunol*. (2020) 11:821. doi: 10.3389/fimmu.2020.00821/full
- Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, et al. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res*. (2016) 44:D73–80. doi: 10.1093/nar/gkv1226
- NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. (2017) 45:D12–7. doi: 10.1093/nar/gkw1071
- Finstermeier K, Zinner D, Brameier M, Meyer M, Kreuz E, Hofreiter M, et al. A mitogenomic phylogeny of living primates. *PLoS One*. (2013) 8:e69504. doi: 10.1371/journal.pone.0069504
- Gordon D, Huddlestone J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, et al. Long-read sequence assembly of the gorilla genome. *Science*. (2016) 352: aae0344. doi: 10.1126/science.aae0344
- O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data*. (2024) 11:732. doi: 10.1038/s41597-024-03571-y
- Xu X, Arnason U. A complete sequence of the mitochondrial genome of the western lowland gorilla. *Mol Biol Evol*. (1996) 13:691–8. doi: 10.1093/oxfordjournals.molbev.a025630
- Mao Y, Harvey WT, Porubsky D, Munson KM, Hoekzema K, Lewis AP, et al. Structurally divergent and recurrently mutated regions of primate genomes. *Cell*. (2024) 187:1547–1562.e13. doi: 10.1016/j.cell.2024.01.052
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. (1990) 215:403–10. doi: 10.1016/S0022-2836(05)80360-2
- Lefranc MP, Lefranc G. IMGT®Homo sapiens IG and TR loci, gene order, CNV and haplotypes: new concepts as a paradigm for jawed vertebrates genome assemblies. *Biomolecules*. (2022) 12:381. doi: 10.3390/biom12030381
- Giudicelli V. IMGT/LIGM-DB, the IMGT(R) comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res*. (2006) 34:D781–4. doi: 10.1093/nar/gkj088
- Lane J, Duroux P, Lefranc MP. From IMGT-ONTOLOGY to IMGT/LIGMOTif: the IMGT® standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC Bioinf*. (2010) 11:223. doi: 10.1186/1471-2105-11-223
- Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*. (2018) 27:135–45. doi: 10.1002/pro.v27.1
- Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, igSF, and mhSF. *Cold Spring Harb Protoc*. (2011) 2011:pdb.ip85. doi: 10.1101/pdb.ip85
- Lefranc MP, Pommier C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol*. (2003) 27:55–77. doi: 10.1016/S0145-305X(02)00039-3
- Giudicelli V, Lefranc MP. IMGT-ONTOLOGY 2012. *Front Gene*. (2012) 3:79. doi: 10.3389/fgene.2012.00079/abstract
- Wong K. Scientific American. Tiny Genetic Differences between Humans and Other Primates Pervade the Genome (2014). Available online at: <https://www.scientificamerican.com/article/tiny-genetic-differences-between-humans-and-other-primates-pervade-the-genome/> (Accessed May, 2022).

38. Giudicelli V. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* (2004) 33:D256–61. doi: 10.1093/nar/gki010
39. Cochrane G, Bates K, Apweiler R, Tateno Y, Mashima J, Kosuge T, et al. Evidence standards in experimental and inferential INSDC third party annotation data. *OMICS: A J Integr Biol.* (2006) 10:105–13. doi: 10.1089/omi.2006.10.105
40. Lefranc MP. From IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures. *Cold Spring Harb Protoc.* (2011) 2011.pdb.ip83. doi: 10.1101/pdb.ip83
41. Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, et al. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res.* (2019) 47:W260–5. doi: 10.1093/nar/gkz303
42. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* (2024) 52(W1):gkae268. doi: 10.1093/nar/gkae268
43. Levinson AI, Kozlowski L, Zheng Y, Wheatley L. B-cell superantigens: definition and potential impact on the immune response. *J Clin Immunol.* (1995) 15:26S–36S. doi: 10.1007/BF01540891
44. Deacy AM, Gan SKE, Derrick JP. Superantigen recognition and interactions: functions, mechanisms and applications. *Front Immunol.* (2021) 12:731845. doi: 10.3389/fimmu.2021.731845
45. Olivieri DN, Gambón Deza F. Immunoglobulin genes in primates. *Mol Immunol.* (2018) 101:353–63. doi: 10.1016/j.molimm.2018.07.020
46. Garzón-Ospina D, Buitrago SP. Immunoglobulin heavy constant gamma gene evolution is modulated by both the divergent and birth-and-death evolutionary models. *Primates.* (2022) 63:611–25. doi: 10.1007/s10329-022-01019-8
47. Barbié V, Lefranc MP. The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp Clin Immunogenet.* (1998) 15:171–83. doi: 10.1159/000019068
48. Wen-Hsiung LL, Dan G. *Fundamentals of Molecular Evolution*. 1st ed. Vol. 252. Sunderland: Sinauer Associates; 1991. 284 p.
49. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. 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
50. Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, et al. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res.* (2011) 21:1626–39. doi: 10.1101/gr.117242.110
51. Attanasio R, Jayashankar L, Engleman CN, Scinicariello F. Baboon immunoglobulin constant region heavy chains: identification of four IGHG genes. *Immunogenetics.* (2002) 54:556–61. doi: 10.1007/s00251-002-0505-1
52. Ueda S, Matsuda F, Honjo T. Multiple recombinational events in primate immunoglobulin epsilon and alpha genes suggest closer relationship of humans to chimpanzees than to gorillas. *J Mol Evol.* (1988) 27:77–83. doi: 10.1007/BF02099732
53. Corpet F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* (1988) 16:10881–90. doi: 10.1093/nar/16.22.10881
54. Ueda S, Takenaka O, Honjo T. A truncated immunoglobulin epsilon pseudogene is found in gorilla and man but not in chimpanzee. *Proc Natl Acad Sci USA.* (1985) 82:3712–5. doi: 10.1073/pnas.82.11.3712
55. Taub RA, Hollis GF, Hieter PA, Korsmeyer S, Waldmann TA, Leder P. Variable amplification of immunoglobulin λ light-chain genes in human populations. *Nature.* (1983) 304:172–4. doi: 10.1038/304172a0
56. Ermert K, Mitlöchner H, Schempp W, Zachau HG. The immunoglobulin κ locus of primates. *Genomics.* (1995) 25:623–9. doi: 10.1016/0888-7543(95)80003-5
57. Brensing-Küppers J, Zocher I, Thiebe R, Zachau HG. The human immunoglobulin κ locus on yeast artificial chromosomes (YACs). *Gene.* (1997) 191:173–81. doi: 10.1016/S0378-1119(97)00056-5
58. Schäble KF, Zachau HG. The variable genes of the human immunoglobulin kappa locus. *Biol Chem Hoppe Seyler.* (1993) 374:1001–22.
59. Lautner-Rieske A, Huber C, Meindl A, Pargent W, Schäble KF, Thiebe R, et al. The human immunoglobulin χ locus. Characterization of the duplicated A regions. *Eur J Immunol.* (1992) 22:1023–9. doi: 10.1002/eji.1830220422
60. Huber C, Huber E, Lautner-Rieske A, Schäble KF, Zachau HG. The human immunoglobulin χ locus. Characterization of the partially duplicated L regions. *Eur J Immunol.* (1993) 23:2860–7. doi: 10.1002/eji.1830231120
61. Yunis JJ, Prakash O. The origin of man: A chromosomal pictorial legacy. *Science.* (1982) 215:1525–30. doi: 10.1126/science.7063861
62. Rayamajhi N, Cheng CHC, Catchen JM. Evaluating Illumina-, Nanopore-, and PacBio-based genome assembly strategies with the bald notothen, *Trematomus borchgrevinkii*. *G3 Genes/Genomes/Genetics.* (2022) 12:jkac192. doi: 10.1093/g3journal/jkac192
63. Tvedte ES, Gasser M, Sparklin BC, Michalski J, Hjelmén CE, Johnston JS, et al. Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. *G3 Genes/Genomes/Genetics.* (2021) 11:jkab083. doi: 10.1093/g3journal/jkab083
64. Yu W, Luo H, Yang J, Zhang S, Jiang H, Zhao X, et al. Comprehensive assessment of 11 *de novo* HiFi assemblers on complex eukaryotic genomes and metagenomes. *Genome Res.* (2024) 34:326–40. doi: 10.1101/gr.278232.123