frontiers in
IMMUNOLOGY

REVIEW ARTICLE
published: 02 September 2013
doi: 10.3389/fimmu.2013.00263

The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor

Katherine J. L. Jackson*, Marie J. Kidd, Yan Wang and Andrew M. Collins

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

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

Reviewed by:
Ramit Mehr, Bar-Ilan University, Israel
Gur Yaari, Yale University, USA
Nir Friedman, Weizmann Institute of Science, Israel

****Correspondence:***
Katherine J. L. Jackson, School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: katherine.jackson@unsw.edu.au

Both the B cell receptor (BCR) and the T cell receptor (TCR) repertoires are generated through essentially identical processes of V(D)J recombination, exonuclease trimming of germline genes, and the random addition of non-template encoded nucleotides. The naïve TCR repertoire is constrained by thymic selection, and TCR repertoire studies have therefore focused strongly on the diversity of MHC-binding complementarity determining region (CDR) CDR3. The process of somatic point mutations has given B cell studies a major focus on variable (IGHV, IGLV, and IGKV) genes. This in turn has influenced how both the naïve and memory BCR repertoires have been studied. Diversity (D) genes are also more easily identified in BCR VDJ rearrangements than in TCR VDJ rearrangements, and this has allowed the processes and elements that contribute to the incredible diversity of the immunoglobulin heavy chain CDR3 to be analyzed in detail. This diversity can be contrasted with that of the light chain where a small number of polypeptide sequences dominate the repertoire. Biases in the use of different germline genes, in gene processing, and in the addition of non-template encoded nucleotides appear to be intrinsic to the recombination process, imparting "shape" to the repertoire of rearranged genes as a result of differences spanning many orders of magnitude in the probabilities that different BCRs will be generated. This may function to increase the precursor frequency of naïve B cells with important specificities, and the likely emergence of such B cell lineages upon antigen exposure is discussed with reference to public and private T cell clonotypes.

Keywords: BCR repertoire, TCR repertoire, V(D)J recombination, public clonotypes, private clonotypes, combinatorial diversity, junctional diversity

GERMLINE GENES AND LYMPHOCYTE DIVERSITY

The mammalian immune system has the ability to respond to almost any antigen to which it is exposed because of the incredible diversity of lymphocyte receptor molecules. The diversity of both the B cell receptor (BCR) repertoire and the T cell receptor (TCR) repertoire is made possible by multiple sets of highly similar genes that recombine to form functional genes. Immunoglobulin heavy chains are encoded by recombined VDJ genes that are formed from sets of Variable (V), Diversity (D), and Joining (J) genes (IGHV, IGHJ, IGHD), while VJ rearrangements of kappa and lambda chain V genes (IGKV, IGLV) and J genes (IGKJ, IGLJ) encode the immunoglobulin light chains (1, 2). TCR β-chains and δ-chains are similarly encoded by distinct sets of V, D, and J genes (TRBV, TRBD, TRBJ; TRDV, TRDD, TRDJ), while α-chains and γ-chains are encoded by additional sets of V and J genes (TRAV, TRAJ; TRGV, TRGJ) (3–5). The resulting combinatorial diversity is expanded still further by junctional diversification arising from exonuclease trimming of the recombining gene ends and from the essentially random addition of nucleotides, between the recombining genes, by the enzyme terminal deoxynucleotidyl transferase (TdT) (6). Together, combinatorial diversity and junctional diversity create the diversity of the naïve T cell and B cell repertoires. Limitations to diversity may however be a feature of V(D)J rearrangement that is as significant to immune function as the bewildering number of lymphocyte specificities that can theoretically be generated.

This review will present evidence that biases in the processes that generate combinatorial and junctional diversity are such that the probabilities of different BCRs and TCRs being generated is highly variable. This results in B and T cells of some specificities being present within the naïve repertoire at high frequency, while other specificities may or may not be present at all. The unevenness of the receptor abundance distribution can be said to give "shape" to the naïve B and T lymphocyte repertoires. This distribution may be further shaped by processes including positive and negative selection, clonal expansion and, in the case of immunoglobulin genes, by somatic hypermutation, however this review will focus upon recombination and gene processing.

As the shape of the naïve human B and T cell lymphocyte repertoire is an outcome of the evolution of genetically determined biases, this should ensure the presence of critical rearrangements in the repertoire of all individuals. It should also ensure that these critical rearrangements are carried by multiple naïve cells (see **Figure 1**). Such populations of specific naïve lymphocytes will have a competitive advantage during antigen-driven clonal selection, and any discussion of repertoire diversity that is limited to the size of the population of unique receptors will therefore be ignoring a parameter of likely biological significance. In this review, we

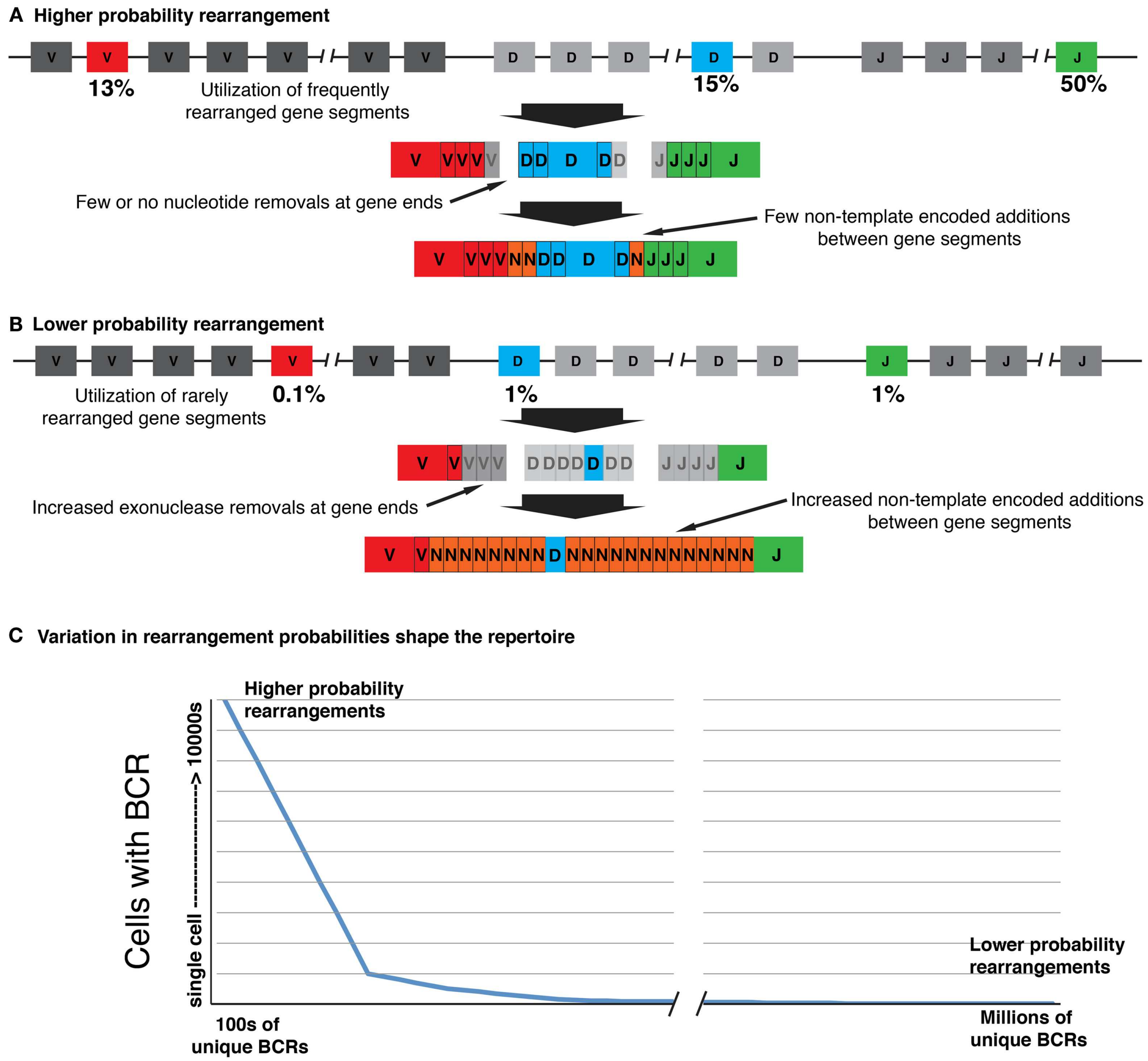

FIGURE 1 | The receptor repertoire has "shape," as biases and constraints in the recombination process vary the probabilities of generating particular V(D)J rearrangements. **(A)** Higher probability rearrangements are generated through utilization of more frequently rearranged gene segments. These segments are joined with minimal gene processing and N-addition, increasing the chance of independent rearrangements with identical or near identical CDR3s. **(B)** Lower probability rearrangements utilize rarely rearranged germline gene segments and the CDR3s are more diverse owing to increased nucleotide removals and additions at the joins. **(C)** The many order of magnitude differences in the likelihood of generating particular rearrangements shape the repertoire, with higher probability rearrangements being frequently generated and as a consequence being carried by many identical B-cells. Only a relatively small number of unique rearrangements will be generated with probabilities high enough to be carried by a large number of B cells, but this should ensure that they are always present in the repertoire at significant levels. Conversely, lower probability rearrangements may be so rare that they are carried only by a single B cell, or are entirely absent from the repertoire. The lower probability rearrangements that are carried by just one or at most a few B-cells likely represent many millions of unique rearrangements.

will use the term "repertoire" to refer to the complete set of receptors that are carried by an individual, including multiple copies of particular sequences. The number of unique sequences that are found within an individual's repertoire will be described as the "diversity" of the repertoire.

The size of the sets of germline genes make a major contribution to lymphocyte diversity, but surprisingly, our knowledge of these germline genes is far from complete. In part this is the result of the complexity of the loci, for they feature numerous highly similar genes that are thought to have evolved via gene conversion (7), and duplication and divergence (8). These genes are interspersed with many pseudogenes and repetitive elements (8). Sequencing and annotation of the loci is therefore challenging. These complexities also mean that SNPs arising from short read-length sequences generated in studies such as the HapMap and 1000 Genomes projects, cannot be used for the imputation of full-length allelic variants. In fact, these projects utilize polyclonal lymphoblastoid cells lines in which the immunoglobulin loci have undergone somatic recombination, and the rearranged genes may have been affected by somatic point mutation. This makes these cell lines unsuited to the study of immunoglobulin genes (9).

Arguably, it is the BCR germline genes that are best known, and paradoxically, this is because of their transformation through the process of somatic hypermutation, during an immune response. IGHV genes are by far the longest of the recombining IGH genes, and they are the principal targets of the mutational machinery (10, 11). Many studies of the immunogenetics of immunoglobulin have therefore concentrated upon the IGHV genes. As it is necessary to be certain of the germline origin of mutated sequences, if accurate studies of point mutations are to be conducted, the complete and accurate definition of the set of germline immunoglobulin IGHV genes and allelic variants has been and should remain a focus of research.

The official human IGHV germline gene dataset, curated by the ImMunoGeneTics (IMGT) group, includes 129 functional genes, open reading frames (ORF), and pseudogenes, as well as over 200 allelic variants (12). Interest in these germline genes has increased in recent years, resulting in 40 new allelic variants being reported since 2005 (13–17). Many additional IGHV allelic variants have also been identified in recent high-throughput sequencing studies, through analysis of cDNA-derived VDJ gene rearrangements (18, 19), but these have not been accepted as part of the official IGHV dataset. We have designated alleles identified in this way with unofficial allele names using an indicator ("p") of their "putative" nature (e.g., IGHV3-9*p03) (15), and these additional alleles can be found in the UNSWIg repertoire (http://www.ihmmune.unsw.edu.au/unswig.php).

The official human light chain V gene datasets appear to be relatively complete and accurate, though few allelic variants have been reported (20). Nevertheless these few variants appear to be of functional and clinical significance. For example, a variant kappa gene allele was identified within the Navajo population and has been reported to account for the susceptibility of this population to infections (21).

The human IGH germline genes receive continuing attention while the IMGT human TCR germline gene datasets have barely changed since the complete sequences of the TCR gene loci were first described (22, 23). The IMGT TRBV dataset includes 65 functional genes, ORFs and pseudogenes, and just 13 allelic variants, and no new TRBV sequence has been added to the dataset since the publication of the complete sequence of the TRB locus in 1996 (22). Only three TRAV/TRDV sequences (24) in the IMGT dataset are derived from studies published since the reporting of the complete sequence of the TRAV/TRDV locus (23), and some variants that were described soon afterward still remain officially unrecognized (25). The incomplete nature of the IMGT TRBV, and TRAV datasets in particular are clearly highlighted in the literature, for sequencing studies have reported many SNPs in the coding regions of these genes. Subramanyan and colleagues reported 279 SNPs in a study of 63 TRBV genes in 10 individuals from each of four human populations (26). Of these reported SNPs, 114 were located in coding regions of functional TRBV genes (26). A similar study of 57 TRAV/TRDV genes in the same 40 individuals resulted in the discovery of 284 SNPs, 51 of which encode amino acid changes in the coding regions of the gene sequences (27). The allelic variants associated with these TRAV/TRDV and TRBV SNPs have not been reported in the literature or in sequence databases, and they have not been incorporated into the official gene datasets. This is surprising because the SNPs were identified through amplification and sequencing of full-length genomic sequences. It is also unfortunate, for studies of TCR polymorphisms have shown that they can be of functional significance (28, 29).

The BCR and TCR D loci contribute differently to the generation of diversity, and the differences in the nature of the loci have influenced BCR and TCR research directions. The 27 human IGHD genes include 25 functional genes, 23 of which are unique (30). Although some IGHD genes, especially those of the IGHD1 gene family, are very similar, there is considerable sequence diversity amongst the genes. The lengths of the IGHD genes vary from 11 nucleotides to 37 nucleotides, and almost all of them are substantially longer than the TRBD and TRDD genes. This length and the IGHD gene variability have made improvement in the identification of IGHD genes within VDJ rearrangements a challenging but achievable research goal. Pursuit of this goal has driven the development of immunoglobulin gene alignment utilities including SODA2 (31), IgBLAST (32), and iHMMune-align (33). The objective measurement of the performance of these utilities is made difficult, however, by a lack of appropriate data sets. Ideally, performance would be measured using rearranged sequences of known composition. As such sets are unavailable, clonally related sequences can be used (32, 33). We have also compared the performance of different utilities using a set of long-read pyrosequenced (Roche 454) IGH rearrangements from an individual with a homozygous deletion of six IGHD genes (34). This test measures performance by the number of VDJ rearrangements in the dataset that are said to include the absent IGHD genes. Together these studies demonstrate that IGHD genes can now be identified with confidence, and as a consequence, analysis of the BCR heavy chain complementarity determining region (CDR) 3 can include detailed analysis of IGHD gene usage, gene processing, and N nucleotide addition.

Analysis of the TCR CDR3 is not so easy. The two human TRBD genes are both short (12 and 16 nucleotides) and highly similar at their 5′ ends (22). This makes their identification in VDJ rearrangements particularly difficult. The TRBD genes within a VDJ rearrangement are likely to be flanked by N-REGIONS of non-template encoded nucleotides. These nucleotides are introduced through the action of the TdT enzyme, which is biased to the addition of guanine (G) nucleotides (35) and to the addition of homopolymer tracts (36, 37). Distinguishing TRBD gene ends from G-rich N nucleotides is difficult because the TRBD genes are G-rich at both their 5′ and 3′ ends. A final complication is that the two alleles of TRBD2 differ by just a single nucleotide. This critical nucleotide is flanked on both sides, in both alleles, by GGG motifs. For these reasons, few TCR studies have included detailed analysis of TRBD genes and their processing, or of the N-REGIONS that can only be defined after the identification of a TRBD gene segment within the CDR3. Even the most recently developed TCR alignment utility excludes identification of TRBD genes from its output (38).

Analysis of the VDJ junction in TRD rearrangements is equally difficult. The three human TRDD genes are just 8, 9, and 13 nucleotides in length (4). This makes their reliable identification within VDJ rearrangements especially problematic if nucleotides have been lost through exonuclease activity. Application of an approach previously used in the analysis of BCR sequences (37) suggests that eight nucleotides is the minimum D gene length that will allow TRDD genes to be reliably distinguished from N-REGIONS within a junction of 12 or fewer nucleotides, while 9 nucleotides are needed for regions from 13 to 15 nucleotides and 10 nucleotides for junctions greater than 15 nucleotides (Jackson, unpublished data). It is therefore no surprise that few studies have reported the partitioning of TRD junctions as two of the three TRDD genes can only be confidently delineated from N-additions in their unprocessed form.

The J loci of the human BCR and TCR also include important differences. The IGHJ locus includes six functional genes, which are all found downstream of the IGHD locus in a single cluster. Allelic variants have been reported for IGHJ3, IGHJ4, IGHJ5, and IGHJ6, though there is reason to doubt the existence of the reported allelic variants of IGHJ3 and IGHJ5 (39). TCR J genes are more numerous and are differently organized. The TRBJ genes are found as a block of six genes located downstream from the TRDB1 gene, and a block of seven genes located downstream from the TRDB2 gene. The TRDB1 gene can pair with all J genes, but the TRDB2 gene is strongly biased toward pairing with its associated J genes (40). There are also four functional J genes in the TRDJ locus. Functional allelic variants have only been reported for the TRBJ1-6 gene.

BIASES IN COMBINATORIAL DIVERSITY AND THE SHAPING OF THE REPERTOIRE

Combinatorial diversity is that part of repertoire diversity that results from the fact that functional receptor genes form by the recombination of members of the sets of germline V, D, and J genes. This diversity is usually calculated by simply multiplying together the number of functional V, D, and J genes that are available within the genome. Such calculations, however, may promote misunderstandings, for they encourage the view that "all genes are equal," and that all combinations are equally likely. TCR studies have paid considerable attention to capturing an unbiased sampling of the repertoire, for example using 5′ RACE to amplify TCR transcripts from the constant region gene. Such studies have shown that TCR genes are highly biased in their usage (41–43). In contrast, many BCR repertoire studies have amplified both mRNA and genomic rearrangements, often using IGHV gene family-targeting primer sets that were developed for the detection of malignancies rather than for the investigation of the repertoire (44, 45). Such primers almost certainly lead to some distortions in the relative abundances of different sequences that are seen. Nevertheless, BCR studies utilizing different primer sets, and amplifying different source material are surprisingly consistent, and the B cell literature provides unequivocal evidence of strong gene utilization biases.

Different IGHV genes are used at frequencies that range from as little as 0.1% to more than 10% of all rearrangements in an individual's naïve B cell repertoire (18, 46). Utilization frequencies also vary between alleles. For example, analysis of VDJ recombination in different individuals has shown that IGHV1-2*02 is used approximately three times as often as IGHV1-2*04, in individuals who carry both these alleles (18). IGHV utilization frequencies are surprisingly constant between individuals (47). Examples of such consistency include IGHV1-46 which varies from 2 to 3.1% in different individuals (average 2.65%), IGHV3-21 which varies from 3.5 to 6.3% (average 4.59%), and IGHV3-49 which varies from 0.8 to 1.3% (average 1.0%) (18). This is not true for all genes, with different individuals utilizing IGHV1-69 at frequencies that range from 3.1 to 9.1% (average 6.2%) (18). IGHV3-23, which is typically the most utilized IGHV gene, was seen on average in 6.7% of all VDJ sequences, but its utilization frequency in one individual was 13.7% (18).

Biased gene usage is not confined to the IGHV genes. IGHD gene usage varies from less than 1% (IGHD4-4/11) to over 15% (IGHD3-22) of total rearrangements. Biases in the resulting amino acid sequences of the CDR3 junction are even greater. IGHD segments can be utilized in all three reading frames, and each IGHD gene is therefore able to encode three distinct amino acid sequences. Analysis of IGH rearrangements in which the IGHJ is out-of-frame, and which are therefore non-productive, shows each IGHD gene rearranges at equal frequency in each of the three RFs, however among productive rearrangements there is a strong skewing of the utilization of each gene toward a dominant RF (48). This dominance is constant between individuals, and the preferred RF is gene family dependent. Analysis of in-frame and out-of-frame IGH rearrangements sequenced using the Illumina platform suggests that the underlying rearrangement processes have no reading frame bias, but that bias emerges from stronger negative selection of sequences in certain reading frames (48). Such negative selection particularly focuses on non-productive sequences that result from the presence of stop codons within the junction region. These are seen when many IGHD genes are translated in the non-dominant reading frame, and such genes can only be utilized in those reading frames if the stop codons are removed by exonuclease trimming. When analysis of IGHD usage in the expressed repertoire factors in the three RFs, the IGHD gene utilization frequencies span three orders of magnitude. There is also considerable variation between the utilization frequencies of IGHJ genes. The IGHJ4 gene is present in approximately 45–50% of rearrangements, while IGHJ6 accounts for a further 20–25% of VDJ rearrangements (49, 50). IGHJ1, on the other hand, is utilized by only 1% of all rearrangements (39).

Biases in light chain gene usage are just as strong. For IGK rearrangements, preferential inclusion of IGKV3-20 was noted in early studies of the expressed IGK repertoire of both adults and neonates (51–53), while single cell PCR (54) and bioinformatics analysis of IGK rearrangements from sequence databases showed IGKV3-15, IGKV3-11, IGKV1-5, IGKV2-30, and IGKV1-30/IGKV1D-39 to also display preferential rearrangement (20). These biases were confirmed again recently in a high-throughput sequencing study which also highlighted similarities in usage between individuals, including similarities between individuals from geographically distant and ethnically distinct populations (55). Under-utilization and over-utilization of the J gene segments have been reported. IGKJ1 and IGKJ2 appear more frequently, while there is under-utilization of IGKJ3 and IGKJ5 (20, 53, 54). This skewing of IGKJ usage toward the genes located 5′ in the IGKJ locus is seen despite the necessity for selection of more 3′ IGKJ genes during secondary IGK rearrangements (56, 57). A similar bias toward 5′ IGKJ genes is also seen in the mouse, and modeling of mouse light chain rearrangement supports the strong underlying tendency toward the initial rearrangement of IGKJ1 or IGKJ2 (58). The IGLV usage is strongly skewed toward a limited number of the functional V segments with 3 of the 30 IGLVs accounting for more than 50% of expressed rearrangements, and with individual IGLV segment frequencies ranging from 0.02 to 27% (59). Only four of the seven IGLJ are considered functional (60). The four IGLJ range from almost 55% utilization in the expressed B cell repertoire for IGLJ7, to just 5.5% for IGLJ1 (61).

Although bias in the reading frame of the IGHD gene is the result of selection, other biases appear to be intrinsic to the recombination process, for when analysis is confined to non-productive rearrangements which carry an out-of-frame J-REGION, preferential gene usage is still seen (48). Such sequences are not subject to positive or negative selection. The same biases have been observed among transcripts generated from transgenic mice that carry a human heavy chain mini-locus (62), while in NOD-scid-IL2Rγ null mice that had been reconstituted with human hematopoietic stem cells, typical patterns of biased usage were seen amongst the expressed light chain genes (63). Recent studies in monozygotic twins show that they share utilization frequencies for both the heavy and light chain genes (46, 63), with correlations in a similar range to replicate biological samples. When one twin was investigated following lymphocyte ablation therapy, the reconstituted repertoire showed the same utilization patterns (46). Unrelated individuals did not share this degree of correlation. The biases in utilization frequencies of different V, D, and J genes therefore appear to be genetically determined, and when acted upon by the recombination machinery, the biases in that process give rise to an individual's distinct repertoire. Repertoire shape is therefore directly linked to the genotype of an individual's immunoglobulin gene loci. This has become even clearer since high-throughput sequencing has allowed analysis to focus upon individual chromosomes.

The large datasets that are now being generated by high-throughput sequencing from single individuals are facilitating analysis of the processes that shape the repertoire, but each dataset still represents a mixture of rearrangements from two independently recombining chromosomes. The fact that V(D)J rearrangement is an intra-chromosomal event, however, means that every V(D)J gene rearrangement provides information about the association of different genes on a chromosome. Any heterozygous locus allows each chromosome to be associated with one or the other allele at that gene locus, and large sets of V(D)J rearrangements can be analyzed to determine all the V, D, and J genes that rearrange on each chromosome. This allows the determination of inferred haplotypes (see **Figure 2**).

In practice, the complete inference of V, D, and J gene haplotypes by the analysis of V(D)J rearrangements is only likely to be possible in the case of the IGH locus. Approximately 40% of individuals are heterozygous at the IGHJ6 locus, and the IGHJ6 gene is present in nearly 25% of all rearrangements. It therefore provides an ideal "anchor-point" from which to haplotype the IGH locus. Using this approach, we recently investigated the IGH locus in nine individuals, and showed that all 18 IGH variable region gene

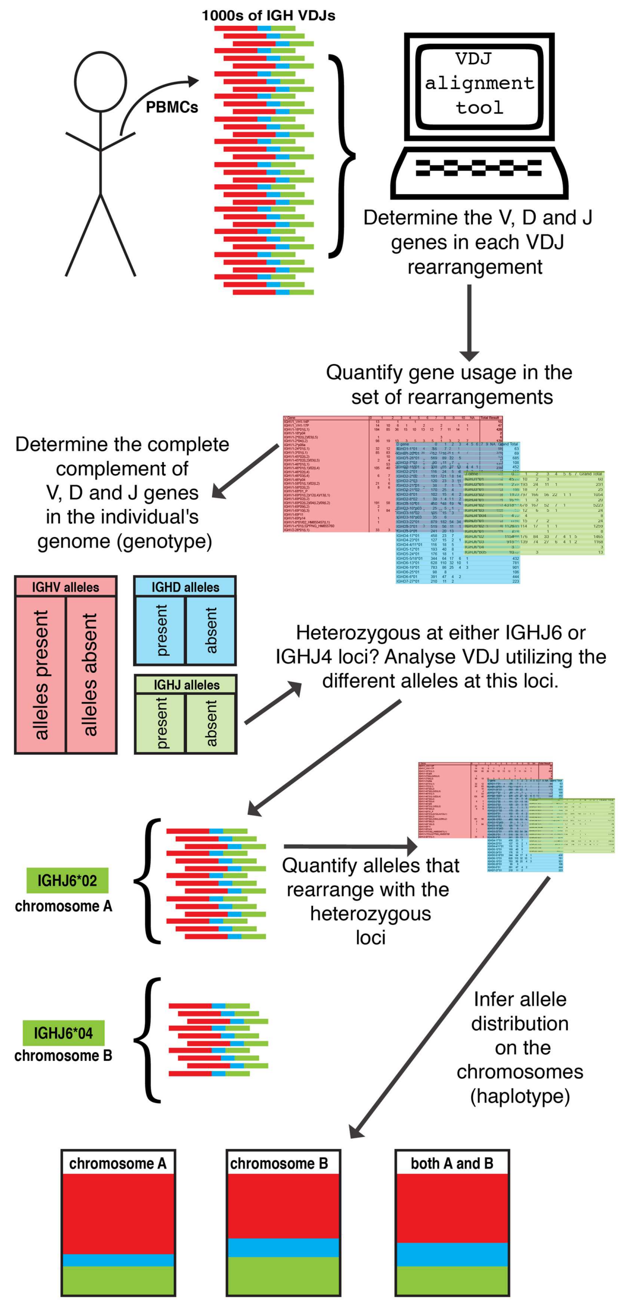

FIGURE 2 | Inference of IGH haplotypes from VDJ rearrangements. The availability of large datasets of VDJ rearrangements from single individuals permit the inference of all germline V, D, and J genes within the genome, from analysis of apparent gene utilization within a dataset. As IGH VDJ rearrangement is an intra-chromosomal event, gene pairing in VDJ rearrangements can be used to infer which gene segments are carried by the same chromosome. Leveraging the heterozygous IGHJ6 or IGHJ4 locus allows the reconstruction of IGH gene segments on each chromosome and for the IGH haplotype to be inferred.

haplotypes were unique (19). In addition to allelic variants, many IGHV and IGHD gene deletions and IGHV gene duplications were evident. The definition of haplotypes in this way is allowing IGH gene usage frequencies to be studied with unprecedented accuracy, but unfortunately no locus as appropriate as the IGHJ6 anchor-point exists amongst the light chain genes or amongst TCR genes. Limited investigations in the past have highlighted TCR haplotypic variation in the human population (64, 65), but the extent of variation within the IGH locus suggests that considerably more TCR variation may await discovery.

Many factors have been explored to explain biases in chromosomal recombination patterns. Variations in enhancers (66) have been implicated in biased murine TCR gene usage. Variations in recombination signal sequences (RSS) also influence utilization frequencies of both human BCR (67, 68) and TCR (69) genes. The IGKV polymorphism that has been linked to increased susceptibility to *Haemophilus influenzae* in the Navajo population includes a single nucleotide change in the heptamer sequence of the RSS, and it reduces recombination by 4.5-fold relative to the common allelic variant (21). The non-amer and heptamer sequences of the RSS are separated by either a 12 or 23 base pair spacer. Spacers also show sequence variation, and there has been debate about the impact this has on recombination efficiency. While some studies did not observe any impact when the regular spacer sequence was replaced with runs of GC pairs (70), competition assays using extra chromosomal substrates suggest differences in spacer sequence can result in differences in recombination efficiency that mirror differential gene usage in the V(D)J repertoire (67, 68). However, RSS variation cannot explain all differences in allele utilization. The recent re-sequencing of the complete IGH locus found that the IGHV-associated RSS were the same as those earlier reported by Matsuda (71) even where different alleles of the gene were present (17).

Some variation in the frequency with which particular gene sequences are seen in the repertoire may be explained by copy-number variations (CNV). The presence of CNV within the IG variable gene locus was first determined using sequence-specific RFLP analysis to determine gene copy-number (72), and the affect of CNV on expression levels was investigated through the examination of the binding of an anti-idiotypic monoclonal antibody (G6) to tonsillar IgD + B-cells (73). An examination of 35 individuals found that they carried between 0 and 4 copies of the IGHV1-69 gene. Linear regression determined that for each allele copy, approximately 3% of B-cells were G6 reactive. Individual differences in the IGHV1-69 copy-number could therefore result in the contribution that this single gene makes varying from being totally absent (0 copies) to being present in as many as 12% of rearrangements in individuals with four available copies.

Sequencing of single chromosomes of an individual's IGH locus has now demonstrated that insertions, deletions, and complex events have altered the copy-number of IGHV genes, including the IGHV1-69 and IGHV3-23 genes (17). The duplicate IGHV3-23 genes remain within the genome as absolutely identical sequences. The presence of these and other CNVs has also been highlighted in bioinformatic studies of immunoglobulin genotypes (18) and haplotypes (19), where sequence data from single individuals clearly demonstrated that some individuals had more than two "alleles" of a single IGHV gene. Genes were also found to be absent from the genome of some individuals. A limitation of these bioinformatics studies was that gene duplications could only be detected if two distinct "allelic variants" were carried on a single chromosome.

In addition to the underlying biases in utilization of germline genes, a final bias has been identified that affects the contribution of recombination frequencies to repertoire diversity. For reasons that are presently unclear, there appear to be pairing preferences for some IGHD and IGHJ genes that increase the frequency of particular IGHD-IGHJ pairs within the repertoire. Biases were first observed in a small set of 59 non-productive rearrangements (74). Later analysis of 6,500 IGH VDJ sequences collected from public databases led to the observation that 5′ IGHD genes paired with increased frequency to the most 3′ IGHJ (J5/J6) and with decreased frequency to the 5′ IGHJ (J1–J4) (50). In contrast, 3′ IGHD tended to preferentially pair with 5′ IGHJ rather than 3′ IGHJ (50). This observation is also supported by analysis of very large datasets generated by pyrosequencing of VDJ rearrangements from three healthy subjects (75). Significantly more pairings were seen of IGHD2-2 and IGHD3-3 with IGHJ6, and of IGHD3-22 and IGHJ3 than would be predicted from the frequencies of these genes in the overall dataset (75).

The bias in D-J pairing also extends to the TCRB loci where the application of HTS approaches to murine TCRB repertoires has revealed a pattern of TRBD to TRBJ pairing that correlates to the genomic distance between rearranged genes (40). The TRBV and TRBJ gene usage in the mice was biased toward particular genes, but the pairings of TRBV and TRBJ were independent. The physical chromatin structure of the TRBD and TRBJ loci was investigated using a biophysical model of the chromatin conformation. The biases in TRBD to TRBJ pairing appeared to be better explained by this mechanical model than previously proposed genetic models based on RSSs (40). The model was also extended to human TRBJ usage with favorable evidence that chromatin conformation determines TRJB gene usage.

Biases in the pairing of heavy and light chains have also been reported. The existence of forbidden or unfavorable pairings of germline heavy and light chain genes was described in the early literature (76). This was not supported by later studies (77, 78), nor was it supported by a recent study that applied high-throughput sequencing to generate thousands of linked heavy and light chain genes (79).

BIASES IN JUNCTIONAL DIVERSITY AND THE SHAPING OF THE REPERTOIRE

Both the naive B cell and T cell repertoires are limited in the periphery by processes of selection. However T cell selection within the thymus is a particularly rigorous process, and it leads to dramatic differences between the potential and the observed repertoire diversity. The idiosyncratic nature of TCR selection in a human population with abundant MHC diversity also means that analysis of the processes that contribute to TCR diversity will be difficult using datasets comprising sequences from multiple individuals. Sufficiently large datasets from single individuals with a specific MHC profile finally became available with the application of high-throughput sequencing to repertoire studies. However, the continuing difficulties involved with the identification of TCR

D genes and hence the other constituent elements within the TCR CDR3 still discourage analysis of the genetic elements and processes that contribute to this region. It is therefore studies of BCR genes that provide the clearest insights into the processes that contribute palindromic P nucleotides and non-template encoded N nucleotides to the V(D)J junction, and into the process of exonuclease trimming that depletes the ends of recombining genes. A recent study of BCR CDR3 suggested that as a result of these processes, the circulating B cell population in a typical adult human includes 3–9 $\times 10^6$ unique heavy chain CDR3 (80).

Palindromic or P nucleotides are formed by the asymmetric opening of hairpin loops that form at gene ends during the rearrangement process (81). In the absence of exonuclease activity, the opening of the hairpins can add short, self-complementary single stranded extensions into the junctions. P nucleotide addition was first recognized as a process that can contribute to TCR CDR3 (82, 83), however the contributions of P nucleotides to the BCR repertoire have been more precisely quantified (84, 85). Similarly, it is recognized that N nucleotides make a major contribution to the diversity of both the TCR and BCR repertoire (86), but only BCR N-REGIONS have been subjected to detailed analysis (37). Where BCR studies have investigated the kinds of amino acids that are likely within N-REGIONS, studies of TCR N-REGIONS have focused upon analysis of the overall contribution of N-REGIONS to αβ TCR diversity. This has been studied in a comparison of wild-type mice with mice carrying homozygous null alleles for TdT (86). N-addition was estimated to contribute to 90% of the diversity of the αβ TCR repertoire (86). Diversity could be estimated in this and other studies because of the development of "spectratyping" techniques, which is the analysis of the CDR3 length distribution in PCR amplicons. It permitted some of the first explorations of the T cell repertoire, however it only allowed detailed analysis of N-REGIONS if further sequencing was undertaken. Until the advent of high-throughput sequencing, such analysis was usually compromised by the restricted number of sequences that could be generated from any individual, and by the challenges associated with D gene identification.

Non-template encoded N-additions are intrinsically biased owing to the preference of TdT toward the incorporation of G nucleotides. This is manifested in G/C-rich additions when viewing the N-REGIONS of the coding strand, as additions may be made to both the coding and non-coding strands during recombination. This has been demonstrated through analysis of extrachromosomal substrates transfected into human cell lines (36), as well as by analysis of human BCR (37) and TCR (87) VDJ rearrangements. The G/C bias is coupled with an apparent interdependence of the additions, which leads to the formation of homopolymer tracts (36, 37, 87). Together these biases ensure that the germline gene-encoded regions of the CDR3 are frequently flanked by amino acids such as glycine, that are encoded by G-rich codons (88). It has been proposed that the inclusion of small amino acids such as glycine, which has only a single side chain, promotes flexibility of the CDR3 loop (88).

Exonuclease trimming is perhaps the least understood process that contributes to the BCR and TCR repertoires. The mechanisms responsible for the loss of nucleotides from the coding ends of the genes during rearrangement remains to be determined, but a number of features of the process have been described, and intrinsic biases have been identified. The extent of processing from each gene end involved in a join (VD or DJ) is independent (87). That is, we do not see more processing on one side of the join to compensate for reduced processing of the gene on the other side. The processing differs for V, D, and J genes and for gene families. Removals may therefore be impacted by the sequence of the gene ends. Sequences with high A/T content appear more susceptible to nucleotide loss, while sequences with high G/C content appear resistant to processing (36, 84, 89, 90). This bias is still seen after controlling for the G/C bias of N-REGIONS.

The gene sequence ends that remain after exonuclease processing provide a final bias that shapes the repertoire. The gene ends are constrained by the genetic code, to favor the formation of codons for a surprisingly limited number of amino acids. This is best illustrated in the case of the many IGHD genes that have the nucleotide sequence TAC at their 3′ end (see **Figure 3**). In the dominant reading frame, these nucleotides encode tyrosine. Removal of a single nucleotide creates a situation where only provision of a T or C (from N-addition or from the 5′ end of the IGHJ gene) will result in a functional sequence, for TAA and TAG are stop codons. Addition of C returns the sequence to its original state, while addition of T results in an alternative tyrosine codon. In this and other cases, the nucleotide sequences of the gene ends limit the diversity that results from exonuclease removals.

B CELL LINEAGES AND T CELL CLONOTYPES IN THE ANTIGEN-SPECIFIC RESPONSE

Biases that we have described in immunoglobulin V, D, and J gene usage mean that at least seven orders of magnitude separate the probabilities that the most likely and the least likely combinations of recombining genes will be generated in the bone marrow. Many additional orders of magnitude separate the most likely from the

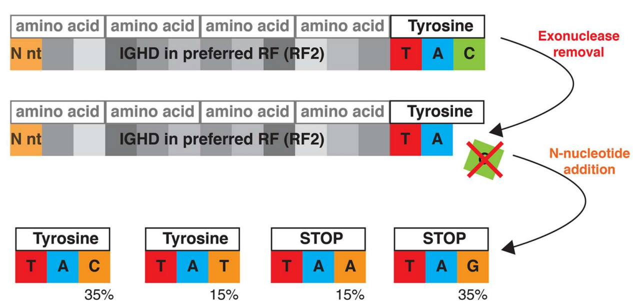

FIGURE 3 | Germline gene segment sequences constrain junctional diversity. This is illustrated by processing of the codon TAC, which encodes tyrosine as the most 3′ amino acid in the preferred reading frame of many IGHD genes. A highly probable outcome of the VDJ rearrangement process is the loss of a single nucleotide from the IGHD coding end. If this occurs, the productive repertoire for these Ds is strongly skewed toward tyrosine, as two of the four possible nucleotide replacement events generate stop codons, and the other two ensure the maintenance of tyrosine. These nucleotides may be sourced from either N-addition or from the 5′ end of an IGHJ segment. The outcome of N-addition is depicted in the bottom panel. N-additions are colored in orange and below each of the four possible outcomes is shown the likelihood of the enzyme TdT adding each nucleotide base.

least likely heavy and light chain pairs. The least likely BCRs are so unlikely to be generated in the bone marrow that they almost certainly will never be seen in an individual's lifetime. The most likely BCRs, on the other hand, may be so readily generated that they are always present within the repertoire at high copy number. These high copy-number sequences are likely to utilize a relative handful of the available germline genes, and to have been subject to minimal processing. TdT adds, on average, around 6 nucleotides between joining genes, and 30 or more nucleotides may occasionally be added to the VD, DJ, and VJ junctions, but it is highly likely that no more than two nucleotides will be added. Even long heavy chain CDR3 are likely to be the result of long germline sequences rather than the result of long N-REGIONS (47). Six or more nucleotides may be removed from the 3′ end of the IGHV gene, but most sequences lose no more than two or three IGHV nucleotides, and many sequences lose no nucleotides at all.

Without the added diversity that comes from D genes, the kappa and lambda repertoires are strongly shaped by biased gene usage and minimal processing and the diversity of the repertoires is surprisingly limited. The light chain repertoires are dominated by a very small number of amino acid sequences, and this dominance is so extreme that even in the days of Sanger sequencing, identical light chain gene rearrangements were reported by separate studies from independent laboratories (20). The theoretical diversity of the kappa repertoire has been estimated to be as high as 4×10^{24} unique nucleotide sequences (91). However analysis of kappa sequences generated from single individuals by high-throughput sequencing suggest the repertoire may include less than 10^4 unique amino acid sequences (55), and some of these sequences may be seen in over 1% of all kappa-bearing BCR (55). The diversity of the expressed lambda repertoire has recently been shown to be similarly restricted (63).

Although the heavy chain repertoire has much greater diversity then the light chain repertoire, repertoire shaping may be sufficiently extreme that some heavy chain sequences, and even some BCR will be present at high copy number in the repertoire of every individual. We are not aware of identical heavy chain sequences being amplified from multiple individuals, but highly similar "stereotypical" sequences have been found amongst leukemic clones of individuals with chronic lymphocytic leukemia (92). These stereotypical sequences differ through the stochastic processes of somatic hypermutation, but they appear to have evolved from cells expressing highly similar BCR within the naïve B cell repertoires of different individuals. Antigen selection, which may be associated with the pathogenesis of this condition (93), could be selecting and therefore revealing high copy-number heavy chain sequences.

The antigen specificity of most heavy and light chain sequences remain unclear, for it is only very recently that antigen-specific human B cells have been isolated and their BCRs investigated. The isolation of antigen-specific plasmablasts from the peripheral blood shortly after vaccination was first used to produce monoclonal human antibodies (94). These cells express BCR genes that are at once similar, as a consequence of their shared origins, yet highly divergent, as a result of the process of somatic point mutation. Together they make up a B cell clone lineage. High-throughput sequencing has since been used to identify clone lineages after booster shots with the influenza vaccine (95) and the pneumococcal vaccine (96). B cell lineages producing broadly neutralizing antibodies to HIV have also been identified using high-throughput sequencing (97). However this handful of studies of antigen-specific B cells in humans has not identified lineages that are shared between individuals. Highly similar BCR heavy chain sequences have recently been identified using high-throughput sequencing of PBMC from multiple individuals with acute symptomatic dengue (98). Although the specificities of these sequences were not determined, such lineages were not identified in uninfected individuals. These may therefore be the first antigen-specific heavy chain "public lineages" to be identified. The extent to which the response to specific antigen more generally involves such "public lineages" remains to be determined.

In contrast to the paucity of studies of antigen-specific B cells, antigen-specific TCRs have been investigated in the human repertoire for over 20 years. Early studies revealed that the immune response to specific antigen, in HLA-matched individuals, can include sets of T cells sharing identical or highly similar TCR α- and β-chains (99–101). The development of techniques for the creation of MHC peptide tetramer complexes has facilitated the identification of antigen-specific T cells by flow cytometry (102). This has allowed the detailed investigation of dominant sequence sets and these studies gave rise to the notions of public and private T cell "clonotypes." Public clonotypes are defined as VDJ amino acid sequences that are dominant and identical, or nearly identical, in multiple individuals. Private clonotypes, in contrast, are idiosyncratic. The apparently antigen-driven emergence of public B cell lineages in chronic lymphocytic B cell leukemia also has parallels amongst T cell leukemias. Studies of T cell large granular lymphocyte leukemias have identified a public clonotype in individuals with the shared DRB1*0701 HLA type (103). This same clonotype was independently identified in DRB1*0701^{+} individuals who were infected with human cytomegalovirus (104), suggesting that antigen-driven pathogenesis may be expanding and revealing this public clonotype.

To understand the reasons for the emergence of particular clonotypes, the naïve repertoire must be better understood. Enrichment techniques have recently been developed which when combined with MHC peptide tetramer technology allows extremely rare peptide-specific naïve murine T cells to be identified (105). Using this approach in humans, naïve CD8^{+} T cells specific for peptide-MHC have been shown to range from 0.6 to 500 cells per million cells (106, 107) and CD4+ T cells to range from 0.2 to 10 per million cells (107). Most of the cells within identified sets of antigen-specific murine T cells express unique TCRs (105), but clonal diversity within identified human cell populations remains unclear. It is likely though that in the much larger human T cell compartment, many circulating T cells could carry identical TCRs. This should ensure that early adaptive responses to these antigens are robust, for the strength of the response to antigen has been shown to reflect the size of the antigen-specific naïve T cell population (105).

The presence of particular public TCR clonotypes have not yet been reported within the naïve human TCR repertoire. Discussion of the emergence of such TCR clonotypes in an antigen-specific response has therefore been driven principally by analyses of their nucleotide and amino acid features, and the phenomenon of convergent recombination has been invoked to explain public clonotypes (108, 109). Many public TCR clonotypes are divergent at the nucleotide level, but identical at the amino acid level. This results from the fact that particular amino acid sequences can arise from multiple, variant nucleotide sequences, and that these nucleotide sequences in turn can sometimes be formed by different genes with varying levels of gene processing and nucleotide addition. Such convergent recombination will certainly contribute to the presence of multiple copies of particular amino acid clonotypes within an individual's repertoire, but arguably, it is unlikely to increase the likelihood of one clonotype over another by more than one or two orders of magnitude.

More recently the role of biases in gene usage and in the recombination process have been identified as an alternative source of public clonotypes (110). The biases in the usage of TCR V, D, and J genes are less pronounced than is the case for the BCR genes. This is the result of the lack of substantial germline diversity within the sets of TRBD and TRDD genes, and because the TRBJ and TRDJ genes lack the strong usage biases that are seen amongst the IGHJ genes. Nevertheless biases in the usage of TCR genes are still likely to ensure that the probabilities of the generation of the least likely and the most likely V(D)J combination seen in αβ and γδ TCR differ by many orders of magnitude. It has also been pointed out that many public clonotypes have short CDR3 loops that are mainly encoded by germline-derived nucleotides rather then TdT-derived nucleotides (110). The contribution this may make to the formation of T cell clonotypes is harder to judge, because of the lack of detailed analysis of these processes, in the context of the TCR repertoire. However lessons from analysis of the BCR repertoire give strong credence to this hypothesis.

Both the BCR and the TCR repertoires have been the subject of considerable study and even greater speculation over many decades. High-throughput sequencing is now revealing their separate secrets at a gratifying rate. Our understanding of the shaping of the BCR and TCR repertoires will now surely move faster if a greater dialog commences between researchers on the two sides of the lymphocyte divide. BCR repertoire studies will be transformed when greater attention is paid to antigen-specific lineages. TCR repertoire studies, in turn, could benefit from the lessons of the BCR repertoire, which suggest that the analysis of full-length V(D)J rearrangements, and detailed analysis of the nucleotide elements within the CDR3, can help explain the shaping of the repertoire.

ACKNOWLEDGMENTS

This work was made possible by a grant from the National Health and Medical Research Council.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) **302**:575–81. doi:10.1038/302575a0
2. Schroeder HW Jr, Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol* (2010) **125**:S41–52. doi:10.1016/j.jaci.2009.09.046
3. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* (1988) **334**:395–402. doi:10.1038/334395a0
4. Takihara Y, Tkachuk D, Michalopoulos E, Champagne E, Reimann J, Minden M, et al. Sequence and organization of the diversity, joining, and constant region genes of the human T-cell delta-chain locus. *Proc Natl Acad Sci U S A* (1988) **85**:6097–101. doi:10.1073/pnas.85.16.6097
5. Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* (2004) **4**:123–32. doi:10.1038/nri1292
6. Desiderio SV, Yancopoulos GD, Paskind M, Thomas E, Boss MA, Landau N, et al. Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature* (1984) **311**:752–5.
7. Davies JM, Platts-Mills TA, Aalberse RC. The enigma of IgE+ B-cell memory in human subjects. *J Allergy Clin Immunol* (2013) **131**:972–6. doi:10.1016/j.jaci.2012.12.1569
8. Fukui K, Noma T, Takeuchi K, Kobayashi N, Hatanaka M, Honjo T. Origin of adult T-cell leukemia virus. Implication for its zoonosis. *Mol Biol Med* (1983) **1**:447–56.
9. 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
10. Rada C, Milstein C. The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. *EMBO J* (2001) **20**:4570–6. doi:10.1093/emboj/20.16.4570
11. Odegard VH, Schatz DG. Targeting of somatic hypermutation. *Nat Rev Immunol* (2006) **6**:573–83. doi:10.1038/nri1896
12. Pallares N, Lefebvre S, Contet V, Matsuda F, Lefranc MP. The human immunoglobulin heavy variable genes. *Exp Clin Immunogenet* (1999) **16**:36–60. doi:10.1159/000019095
13. Ohm-Laursen L, Larsen SR, Barington T. Identification of two new alleles, IGHV3-23*04 and IGHJ6*04, and the complete sequence of the IGHV3-h pseudogene in the human immunoglobulin locus and their prevalences in Danish Caucasians. *Immunogenetics* (2005) **57**:621–7. doi:10.1007/s00251-005-0035-8
14. Romo-Gonzalez T, Morales-Montor J, Rodriguez-Dorantes M, Vargas-Madrazo E. Novel substitution polymorphisms of human immunoglobulin VH genes in Mexicans. *Hum Immunol* (2005) **66**:732–40. doi:10.1016/j.humimm.2005.03.002
15. Wang Y, Jackson KJ, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol* (2008) **86**:111–5.
16. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA, et al. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* (2011) **63**:259–65. doi:10.1007/s00251-010-0510-8
17. 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
18. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. 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
19. 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
20. Collins AM, Wang Y, Singh V, Yu P, Jackson KJ, Sewell WA. The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate. *Immunogenetics* (2008) **60**:669–76. doi:10.1007/s00251-008-0325-z
21. 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
22. Rowen L, Koop BF, Hood L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* (1996) **272**:1755–62. doi:10.1126/science.272.5269.1755

23. Boysen C, Simon MI, Hood L. Analysis of the 1.1-Mb human alpha/delta T-cell receptor locus with bacterial artificial chromosome clones. *Genome Res* (1997) **7**:330–8.
24. Ibberson MR, Copier JP, Llop E, Navarrete C, Hill AV, Cruickshank JK, et al. T-cell receptor variable alpha (TCRAV) polymorphisms in European, Chinese, South American, AfroCaribbean, and Gambian populations. *Immunogenetics* (1998) **47**:124–30. doi:10.1007/s002510050337
25. Moody AM, Reyburn H, Willcox N, Newsom-Davis J. New polymorphism of the human T-cell receptor AV28S1 gene segment. *Immunogenetics* (1998) **48**:62–4. doi:10.1007/s002510050401
26. Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am J Hum Genet* (2001) **69**:381–95. doi:10.1086/321297
27. Mackelprang R, Livingston RJ, Eberle MA, Carlson CS, Yi Q, Akey JM, et al. Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. *Hum Genet* (2006) **119**:255–66. doi:10.1007/s00439-005-0111-z
28. Vessey SJ, Bell JI, Jakobsen BK. A functionally significant allelic polymorphism in a T cell receptor V beta gene segment. *Eur J Immunol* (1996) **26**:1660–3. doi:10.1002/eji.1830260739
29. Gras S, Chen Z, Miles JJ, Liu YC, Bell MJ, Sullivan LC, et al. Allelic polymorphism in the T cell receptor and its impact on immune responses. *J Exp Med* (2010) **207**:1555–67. doi:10.1084/jem.20100603
30. Corbett SJ, Tomlinson IM, Sonnhammer EL, Buck D, Winter G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, “minor” D segments or D-D recombination. *J Mol Biol* (1997) **270**:587–97. doi:10.1006/jmbi.1997.1141
31. Munshaw S, Kepler TB. SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* (2010) **26**:867–72. doi:10.1093/bioinformatics/btq056
32. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013). doi:10.1093/nar/gkt382
33. Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* (2007) **23**:1580–7. doi:10.1093/bioinformatics/btm147
34. Jackson KJ, Boyd S, Gaeta BA, Collins AM. Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset. *Bioinformatics* (2010) **26**:3129–30. doi:10.1093/bioinformatics/btq604
35. Basu M, Hegde MV, Modak MJ. Synthesis of compositionally unique DNA by terminal deoxynucleotidyl transferase. *Biochem Biophys Res Commun* (1983) **111**:1105–12. doi:10.1016/0006-291X(83)91413-4
36. Gauss GH, Lieber MR. Mechanistic constraints on diversity in human V(D)J recombination. *Mol Cell Biol* (1996) **16**:258–69.
37. Jackson KJ, Gaeta BA, Collins AM. Identifying highly mutated IGHD genes in the junctions of rearranged human immunoglobulin heavy chain genes. *J Immunol Methods* (2007) **324**:26–37. doi:10.1016/j.jim.2007.04.011
38. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* (2013) **29**:542–50. doi:10.1093/bioinformatics/btt004
39. Lee CE, Jackson KJ, Sewell WA, Collins AM. Use of IGHJ and IGHD gene mutations in analysis of immunoglobulin sequences for the prognosis of chronic lymphocytic leukemia. *Leuk Res* (2007) **31**:1247–52. doi:10.1016/j.leukres.2006.10.013
40. Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, et al. Chromatin conformation governs T-cell receptor Jbeta gene segment usage. *Proc Natl Acad Sci U S A* (2012) **109**:15865–70. doi:10.1073/pnas.1203916109
41. Quiros Roldan E, Sottini A, Bettinardi A, Albertini A, Imberti L, Primi D. Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells. *Immunogenetics* (1995) **41**:91–100.
42. Livak F, Burtrum DB, Rowen L, Schatz DG, Petrie HT. Genetic modulation of T cell receptor gene segment usage during somatic recombination. *J Exp Med* (2000) **192**:1191–6. doi:10.1084/jem.192.8.1191
43. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) **21**:790–7. doi:10.1101/gr.115428.110
44. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* (2003) **17**:2257–317. doi:10.1038/sj.leu.2403202
45. Matthews C, Catherwood M, Morris TC, Alexander HD. Routine analysis of IgVH mutational status in CLL patients using BIOMED-2 standardized primers and protocols. *Leuk Lymphoma* (2004) **45**:1899–904. doi:10.1080/10428190410001710812
46. 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 U S A* (2011) **108**:20066–71. doi:10.1073/pnas.1107498108
47. Briney BS, Willis JR, Mckinney BA, Crowe JE Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. *Genes Immun* (2012) **13**:469–73. doi:10.1038/gene.2012.20
48. Benichou J, Glanville J, Prak ET, Azran R, Kuo TC, Pons J, et al. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol* (2013) **190**:5567–77. doi:10.4049/jimmunol.1201929
49. Yamada M, Wasserman R, Reichard BA, Shane S, Caton AJ, Rovera G. Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes. *J Exp Med* (1991) **173**:395–407. doi:10.1084/jem.173.2.395
50. Volpe JM, Kepler TB. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res* (2008) **4**:3. doi:10.1186/1745-7580-4-3
51. Klein R, Jaenichen R, Zachau HG. Expressed human immunoglobulin kappa genes and their hypermutation. *Eur J Immunol* (1993) **23**:3248–62. doi:10.1002/eji.1830231231
52. Cox JP, Tomlinson IM, Winter G. A directory of human germ-line V kappa segments reveals a strong bias in their usage. *Eur J Immunol* (1994) **24**:827–36. doi:10.1002/eji.1830240409
53. Weber JC, Blaison G, Martin T, Knapp AM, Pasquali JL. Evidence that the V kappa III gene usage is nonstochastic in both adult and newborn peripheral B cells and that peripheral CD5+ adult B cells are oligoclonal. *J Clin Invest* (1994) **93**:2093–105. doi:10.1172/JCI117204
54. Foster SJ, Brezinschek HP, Brezinschek RI, Lipsky PE. Molecular mechanisms and selective influences that shape the kappa gene repertoire of IgM+ B cells. *J Clin Invest* (1997) **99**:1614–27. doi:10.1172/JCI119324
55. Jackson KJ, Wang Y, Gaeta BA, Pomat W, Siba P, Rimmer J, et al. Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* (2012) **64**:3–14. doi:10.1007/s00251-011-0559-z
56. Gay D, Saunders T, Camper S, Weigert M. Receptor editing: an approach by autoreactive B cells to escape tolerance. *J Exp Med* (1993) **177**:999–1008. doi:10.1084/jem.177.4.999
57. Tiegs SL, Russell DM, Nemazee D. Receptor editing in self-reactive bone marrow B cells. *J Exp Med* (1993) **177**:1009–20. doi:10.1084/jem.177.4.1009
58. Mehr R, Shannon M, Litwin S. Models for antigen receptor gene rearrangement. I. Biased receptor editing in B cells: implications for allelic exclusion. *J Immunol* (1999) **163**:1793–8.
59. Ignatovich O, Tomlinson IM, Jones PT, Winter G. The creation of diversity in the human immunoglobulin V(lambda) repertoire. *J*

Mol Biol (1997) **268**:69–77. doi:10.1006/jmbi.1997.0956

60. Vasicek TJ, Leder P. Structure and expression of the human immunoglobulin lambda genes. *J Exp Med* (1990) **172**:609–20. doi:10.1084/jem.172.2.609
61. Farner NL, Dorner T, Lipsky PE. Molecular mechanisms and selection influence the generation of the human V lambda J lambda repertoire. *J Immunol* (1999) **162**:2137–45.
62. Tuaillon N, Taylor LD, Lonberg N, Tucker PW, Capra JD. Human immunoglobulin heavy-chain minilocus recombination in transgenic mice: gene-segment use in mu and gamma transcripts *Proc Natl Acad Sci U S A* (1993) **90**:3720–4. doi:10.1073/pnas.90.8.3720
63. Hoi KH, Ippolito GC. Intrinsic bias and public rearrangements in the human immunoglobulin Vlambda light chain repertoire. *Genes Immun* (2013) **14**:271–6. doi:10.1038/gene.2013.10
64. Craddock TP, Zumla AM, Ollier WE, Chintu CZ, Muyinda GP, Lancaster FC, et al. Predominance of one T-cell antigen receptor BV haplotype in African populations. *Immunogenetics* (2000) **51**:231–7. doi:10.1007/s002510050036
65. Donaldson IJ, Shefta J, Lawson CA, Bushnell JR, Morgan AW, Isaacs JD, et al. Unique TCR beta-subunit variable gene haplotypes in Africans. *Immunogenetics* (2002) **53**:884–93. doi:10.1007/s00251-001-0406-8
66. McMurry MT, Hernandez-Munain C, Lauzurica P, Krangel MS. Enhancer control of local accessibility to V(D)J recombinase. *Mol Cell Biol* (1997) **17**:4553–61.
67. Nadel B, Tang A, Escuro G, Lugo G, Feeney AJ. Sequence of the spacer in the recombination signal sequence affects V(D)J rearrangement frequency and correlates with nonrandom Vkappa usage in vivo. *J Exp Med* (1998) **187**:1495–503. doi:10.1084/jem.187.9.1495
68. Feeney AJ, Tang A, Ogwaro KM. B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. *Immunol Rev* (2000) **175**:59–69. doi:10.1111/j.1600-065X.2000.imr017508.x
69. Posnett DN, Vissinga CS, Pambuccian C, Wei S, Robinson MA, Kostyu D, et al. Level of human TCRBV3S1 (V beta 3) expression correlates with allelic polymorphism in the spacer region of the recombination signal sequence. *J Exp Med* (1994) **179**:1707–11. doi:10.1084/jem.179.5.1707
70. Wei Z, Lieber MR. Lymphoid V(D)J recombination. Functional analysis of the spacer sequence within the recombination signal. *J Biol Chem* (1993) **268**:3180–3.
71. Matsuda F, Ishii K, Bourvagnet P, Kuma K, 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
72. Sasso EH, Willems Van Dijk K, Bull A, Van Der Maarel SM, Milner EC. VH genes in tandem array comprise a repeated germline motif. *J Immunol* (1992) **149**:1230–6.
73. 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
74. Souto-Carneiro MM, Longo NS, Russ DE, Sun HW, Lipsky PE. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J Immunol* (2004) **172**:6790–802.
75. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* (2009) **1**:12ra23. doi:10.1126/scitranslmed.3000540
76. De Preval C, Fougereau M. Specific interaction between VH and VL regions of human monoclonal immunoglobulins. *J Mol Biol* (1976) **102**:657–78. doi:10.1016/0022-2836(76)90340-5
77. Brezinschek HP, Foster SJ, Dorner T, Brezinschek RI, Lipsky PE. Pairing of variable heavy and variable kappa chains in individual naive and memory B cells. *J Immunol* (1998) **160**:4762–7.
78. de Wildt RM, Hoet RM, Van Venrooij WJ, Tomlinson IM, Winter G. Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. *J Mol Biol* (1999) **285**:895–901. doi:10.1006/jmbi.1998.2396
79. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) **31**:166–9. doi:10.1038/nbt.2492
80. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE* (2011) **6**:e22365. doi:10.1371/journal.pone.0022365
81. Roth DB, Menetski JP, Nakajima PB, Bosma MJ, Gellert M. V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. *Cell* (1992) **70**:983–91. doi:10.1016/0092-8674(92)90248-B
82. Lafaille JJ, Decloux A, Bonneville M, Takagaki Y, Tonegawa S. Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. *Cell* (1989) **59**:859–70. doi:10.1016/0092-8674(89)90609-0
83. Tian C, Luskin GK, Dischert KM, Higginbotham JN, Shepherd BE, Crowe JE Jr. Evidence for preferential Ig gene usage and differential TdT and exonuclease activities in human naive and memory B cells. *Mol Immunol* (2007) **44**:2173–83. doi:10.1016/j.molimm.2006.11.020
84. Jackson KJ, Gaeta B, Sewell W, Collins AM. Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunol* (2004) **5**:19. doi:10.1186/1471-2172-5-19
85. Lu H, Schwarz K, Lieber MR. Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res* (2007) **35**:6917–23. doi:10.1093/nar/gkm823
86. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* (2001) **194**:1385–90. doi:10.1084/jem.194.9.1385
87. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* (2012) **109**:16161–6. doi:10.1073/pnas.1212755109
88. Hofle M, Linthicum DS, Ioerger T. Analysis of diversity of nucleotide and amino acid distributions in the VD and DJ joining regions in Ig heavy chains. *Mol Immunol* (2000) **37**:827–35. doi:10.1016/S0161-5890(00)00110-3
89. Boubnov NV, Wills ZP, Weaver DT. V(D)J recombination coding junction formation without DNA homology: processing of coding termini. *Mol Cell Biol* (1993) **13**:6957–68.
90. Nadel B, Feeney AJ. Influence of coding-end sequence on coding-end processing in V(D)J recombination. *J Immunol* (1995) **155**:4322–9.
91. Saada R, Weinberger M, Shahaf G, Mehr R. Models for antigen receptor gene rearrangement: CDR3 length. *Immunol Cell Biol* (2007) **85**:323–32. doi:10.1038/sj.icb.7100055
92. Agathangelidis A, Darzentas N, Hadzidimitriou A, Brochet X, Murray F, Yan XJ, et al. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood* (2012) **119**:4467–75. doi:10.1182/blood-2011-11-393694
93. Hoogeboom R, Van Kessel KP, Hochstenbach F, Wormhoudt TA, Reinten RJ, Wagner K, et al. A mutated B cell chronic lymphocytic leukemia subset that recognizes and responds to fungi. *J Exp Med* (2013) **210**:59–70. doi:10.1084/jem.20121801
94. Smith K, Garman L, Wrammert J, Zheng NY, Capra JD, Ahmed R, et al. Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat Protoc* (2009) **4**:372–84. doi:10.1038/nprot.2009.3
95. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* (2013) **5**:171ra119. doi:10.1126/scitranslmed.3004794
96. Wu YC, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front Immunol* (2012) **3**:193. doi:10.3389/fimmu.2012.00193
97. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* (2011) **333**:1593–602. doi:10.1126/science.1207532

98. Parameswaran P, Yi Liu Q, Roskin KM, Jackson KK, Dixit VP, Lee J-Y, et al. Coherent immune repertoire signatures in human dengue. *Cell Host Microbe* (Forthcoming 2013).
99. Moss PA, Moots RJ, Rosenberg WM, Rowland-Jones SJ, Bodmer HC, Mcmichael AJ, et al. Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. *Proc Natl Acad Sci U S A* (1991) **88**:8987–90. doi:10.1073/pnas.88.20.8987
100. Argaet VP, Schmidt CW, Burrows SR, Silins SL, Kurilla MG, Doolan DL, et al. Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus. *J Exp Med* (1994) **180**:2335–40. doi:10.1084/jem.180.6.2335
101. Pannetier C, Even J, Kourilsky P. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol Today* (1995) **16**:176–81. doi:10.1016/0167-5699(95)80117-0
102. Altman JD, Moss PA, Goulder PJ, Barouch DH, Mcheyzer-Williams MG, Bell JI, et al. Phenotypic analysis of antigen-specific T lymphocytes. *Science* (1996) **274**:94–6. doi:10.1126/science.274.5284.94
103. Garrido P, Ruiz-Cabello F, Barcena P, Sandberg Y, Canton J, Lima M, et al. Monoclonal TCR-Vbeta13.1+/CD4+/NKa+/CD8-/+dim T-LGL lymphocytosis: evidence for an antigen-driven chronic T-cell stimulation origin. *Blood* (2007) **109**:4890–8. doi:10.1182/blood-2006-05-022277
104. Crompton L, Khan N, Khanna R, Nayak L, Moss PA. CD4+ T cells specific for glycoprotein B from cytomegalovirus exhibit extreme conservation of T-cell receptor usage between different individuals. *Blood* (2008) **111**:2053–61. doi:10.1182/blood-2007-04-079863
105. Moon JJ, Chu HH, Pepper M, Mcsorley SJ, Jameson SC, Kedl RM, et al. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* (2007) **27**:203–13. doi:10.1016/j.immuni.2007.07.007
106. Alanio C, Lemaitre F, Law HK, Hasan M, Albert ML. Enumeration of human antigen-specific naive CD8+ T cells reveals conserved precursor frequencies. *Blood* (2010) **115**:3718–25. doi:10.1182/blood-2009-10-251124
107. Kwok WW, Tan V, Gillette L, Littell CT, Soltis MA, Lafond RB, et al. Frequency of epitope-specific naive CD4(+) T cells correlates with immunodominance in the human memory repertoire. *J Immunol* (2012) **188**:2537–44. doi:10.4049/jimmunol.1102190
108. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci USA* (2006) **103**:18691–6. doi:10.1073/pnas.0608907103
109. Quigley MF, Greenaway HY, Venturi V, Lindsay R, Quinn KM, Seder RA, et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc Natl Acad Sci USA* (2010) **107**:19414–9. doi:10.1073/pnas.1010586107
110. Miles JJ, Douek DC, Price DA. Bias in the alphabeta T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol* (2011) **89**:375–87. doi:10.1038/icb.2010.139

Conflict of Interest Statement: 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.

Received: 31 May 2013; accepted: 19 August 2013; published online: 02 September 2013.
*Citation: Jackson KJL, Kidd MJ, Wang Y and Collins AM (2013) The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. Front. Immunol. **4**:263. doi: 10.3389/fimmu.2013.00263*
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.