2
3
4 **Inter- and intra-species comparison of phylogenetic fingerprints and sequence**
5 **diversity of immunoglobulin variable genes**

6
7 **Alexander Yermanos^{1,2,*}, Andreas Dounas³, Victor Greiff⁴, Tanja Stadler¹,**
8 **Annette Oxenius², Sai T. Reddy^{1,*}**

9
10 ¹Department of Biosystems Science and Engineering, ETH Zurich, Basel,
11 Switzerland

12 ²Institute of Microbiology, ETH Zurich, Zurich, Switzerland,

13 ³Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich,
14 Switzerland

15 ⁴Department of Immunology, University of Oslo, Oslo, Norway

16 *Correspondence: ayermanos@gmail.com, sai.reddy@bsse.ethz.ch

17
18 **Abstract**

19 Protection and neutralization of a vast array of pathogens is accomplished by the
20 tremendous diversity of the B cell receptor (BCR) repertoire. For jawed vertebrates,
21 this diversity is initiated via the somatic recombination of immunoglobulin (Ig)
22 germline elements. While it is clear that the number of these germline segments differs
23 from species to species, the extent of cross-species sequence diversity remains largely
24 uncharacterized. Here we use extensive computational and statistical methods to
25 investigate the sequence diversity and evolutionary relationship between Ig variable
26 (V), diversity (D), and joining (J) germline segments across nine commonly studied
27 species ranging from zebrafish to human. Metrics such as guanine-cytosine (GC)
28 content showed low redundancy across Ig germline genes within a given species. Other
29 comparisons, including amino acid motifs, evolutionary selection, and sequence
30 diversity revealed species-specific properties. Additionally, we showed that the
31 germline-encoded diversity differs across antibody (recombined VDJ) repertoires of
32 various B cell subsets. To facilitate future comparative immunogenomics analysis we
33 created VDJgermlines, an R package that contains the germline sequences from
34 multiple species. Our study informs strategies for the humanization and engineering of
35 therapeutic antibodies.

Introduction

B cells are capable of interacting with a seemingly infinite array of foreign antigens via the binding of their characteristic BCR (or secreted version: antibodies). The theoretical diversity of unique BCRs for most species is often orders of magnitudes higher than the actual physiological number of B cells present in a given individual (Elhanati et al. 2014; Greiff et al. 2017a). Across jawed vertebrates, this tremendous diversity is accomplished primarily in two manners. The first step involves a form of somatic recombination, in which germline gene segments recombine to form the variable regions of the BCR. The variable region of the heavy chain arises from the recombination of variable– (V), diversity– (D), and joining– (J) germline segments, whereas the light chain recombines using either a variable kappa (V_{κ}) or variable lambda (V_{λ}) segment with an accompanying J-segment (Tonegawa 1983). During this recombination, imprecise junctional joining occurs, and nucleotides are randomly inserted or deleted at the junctions, providing further sequence diversity (Nadel and Feeney 1997; Glanville et al. 2009; Elhanati et al. 2015). The second step of diversification occurs on rearranged V(D)J regions via mechanisms such as somatic hypermutation or gene conversion (Methot and Di Noia 2017; Weber et al. 2017). The increased availability of high-throughput sequencing (HTS) data has provided a wealth of information regarding the genetic nature of the humoral immune response. Specifically, advancements in sequencing technologies have elucidated both the number and sequences of germline-encoded immunoglobulin constituents in multiple species (Schiaffella et al. 1999; Johnston et al. 2006; Howe et al. 2013), which has led to the creation of specialized databases containing this information (Lefranc et al. 1999; Lees et al. 2019). Further tools have been developed to predict germline genes from whole shotgun sequencing data (Olivieri et al. 2014). Using this germline information, protocols have been developed for targeted HTS of immunoglobulin repertoires (Ig-Seq) by designing gene-specific primers to these regions that enable amplification of entire recombined V-(D)-J regions (Georgiou et al. 2014; Greiff et al. 2015; Miho et al. 2018). The recovered sequencing information provides a quantitative description of the BCR landscape, and has been implemented extensively in studies related to vaccine profiling, diagnostics, and monoclonal antibody discovery (Greiff et al.; Reddy et al. 2010; Vollmers et al. 2013; Laserson et al. 2014; Abbott et al. 2018; Parola et al. 2018). Additionally, multiple bioinformatic tools exist to infer the germline genes from Ig-Seq

experiments (Gadala-Maria et al. 2015; Ralph and Iv 2016; Corcoran et al. 2016; Gidoni et al. 2019).

The promise of obtaining such a comprehensive and quantitative analysis from Ig-seq has driven an explosion of studies implementing this methodology across multiple species (Weinstein et al. 2009; Lavinder et al. 2014; Stanfield et al. 2016; Yermanos et al. 2017; Friedensohn et al. 2018). While both the combinatorial nature of which Ig genes segments recombine and the nature of somatic hypermutation have been extensively studied, less attention has been paid to the underlying somatic diversity of the germline segments themselves. Theoretical diversity arising solely from somatic recombination is traditionally based upon the product of the number of V-, D-, and J- genes, which ignores the extent of sequence diversity and similarity across the segments. For example, the mouse genome contains approximately 70 more immunoglobulin heavy (IgH) V gene segments than the human (Lefranc et al. 1999), however, the increase in potential diversity may be overstated if these murine V-genes are extremely similar to each other.

Evolutionarily, the immunoglobulin genes are assumed to have arisen due to multiple rounds of duplications and diversification events (Janeway and Murphy 2011). While each species faces unique selective pressures during this proposed birth-death process of germline segments, it remains unknown how similar this process was for various species. Furthermore, the extent of this germline diversification (e.g. how different the variable Ig genes are within a species) may provide clues regarding how naïve B cells can recognize a vast number of foreign pathogens. Therefore, we leveraged multiple phylogenetic tools and bioinformatics analyses to better characterize both the sequence diversification and the evolutionary landscape of variable heavy and light germline segments in nine species commonly studied in immunoglobulin research. Furthermore, we included two sets of human IgH V alleles from the Immunogenetics (IMGT) reference database (Lefranc et al. 1999) to simultaneously compare both intra-and inter-species diversity. We hypothesized that the variable (heavy, lambda, and kappa) genes in all species would have experienced convergent selection to maximize sequence diversity, independent of species. We discovered that prominent species-specific effects, such as amino acid motifs or selective pressures, have shaped the genetic landscape of these variable gene segments. We further investigated whether differences exist in variable gene diversity across different populations of murine B cells (pre-B cells, naïve B cells or plasma cells). We found that the sequence diversity of the highly

expressed IgH V segments is specific to the B cell compartment, based on 20 previously reported murine Ig-Seq experiments (Greiff et al. 2017a), despite equal numbers of unique V-genes compared. Furthermore, we developed an R package, VDJgermlines, which contains the germline genes of commonly studied species and provides a user-friendly interface to easily extract nucleotide and amino acid sequences for known BCR gene segments and alleles (in addition to TCR alleles) from multiple databases. The user can specify species, chain, database, and strain to easily and instantaneously extract sequences and accession numbers for multiple sequences simultaneously. These species-specific findings have potential applications in monoclonal antibody design (Brown et al. 2019), for example, in the humanization of antibodies discovered in animal models.

Results

Low redundancy GC content, predicted solubility and protein stability for heavy variable segments

We hypothesized that within a given species, the germline gene segments would have evolved to maximize the germline-encoded diversity. While we have included the kappa and lambda chains in some analyses, we have primarily focused on the V-segments of the heavy chain due to the high number, long sequence length, and importance in dictating antigen specificity (Xu and Davis 2000) (Figures 1, S1A, S1B, S1C). We first compared the GC content across the IgH V germline segments for nine species (human, macaque, mouse, rat, platypus, zebrafish, rabbit, alpaca, and dog) commonly used in antibody or immunoglobulin genetics research (Figure 1). We additionally included two sets of IgH V alleles (either *01 or *02 from the IMGT reference database (Lefranc et al. 1999)). Within the 63 human IgH V genes analyzed, there were 50 unique allelic nucleotide variants between the two reference human allele sets (Figure 1B). Transcript analysis studies have shown that higher GC content can lead to increased steady state mRNA levels (Kudla et al. 2006), which one may expect to remain relatively conserved upon duplication events. If a convergent selective pressure to maximize germline diversity existed within all species, we would expect broad distributions of GC content. While the mean GC content ranged between 44-57% for the various species, we observed low redundancy of GC content for the heavy chains of all nine species analyzed (Figure 1C), despite dramatically different numbers of V-genes per species (Figure S1A). This variation of GC content is largely due to mutation

and not simply increased sequence length, suggested by the highly conserved sequence length of the majority of IgH V genes across each species (Figure S1B). The GC content between the two human allelic variant sets was highly correlated (Figure 1D), suggesting that the aforementioned IgH V diversification is maintained across individuals within a species, as well as across species. Thus, while the mean GC content may differ for each species, the GC content for all V-genes of each species is distributed across a range of 10%, suggesting a similar selective pressure common to multiple species. Interestingly, those species with the most IgH V genes (mouse and rat, Figure S1A), also show the largest range of GC content (40-55%). Furthermore, the distributions of V gene GC content overlap in all species excluding the zebrafish (ranging from ~37-47%), which is also the species separated by the largest evolutionary distance (Figure 2b). These findings may indicate an evolutionary advantage to diversifying the GC content of multiple IgH V (in addition to IgH D and IgH J, Figure S3) genes, which potentially maintains a broad spectrum of steady state mRNA levels. If an evolutionary pressure for such diversification exists, we would expect pseudogenes to maintain a narrower distribution of GC content than those functional V genes. Indeed, further investigation revealed that functional V gene maintained a broader distribution of GC compared to the pseudogenes.

The aforementioned low redundancy of GC content across all included species led us to ask whether this trend remained for metrics based on amino acid sequence properties. We therefore utilized commonly used bioinformatic tools to quantify biophysical characteristics including solubility, aliphatic index, and protein instability on the germline genes (Osorio, D., Rondon-Villarreal, P. & Torres, R 2015; Sormanni et al. 2015, 2017). All three metrics predicted a similar trend of diverse biophysical properties, despite conserved sequence lengths (Figures 1E, 1F, S1E, S1F), further suggesting that the range of immunoglobulin diversity overlaps across different species. We finally asked how similar the inferred biophysical properties of allelic variants were for identical IgH V genes (e.g. IgH V1-1*01 and IgH V1-1*02). Despite the two human allelic sets containing unique sequences for 50 of the 63 IgH V genes, both the solubility and aliphatic indices were highly correlated when comparing the same germline segment (Figures 1G, 1H).

Species-specific germline gene diversification identified by sequence motifs

We next asked whether the sequence compositions of the heavy, lambda, and kappa variable Ig segments showed a similar lack of redundancy within a species as previously observed with GC content, solubility, and protein instability (Figures 1A, 1E, 1F). To answer this, we first quantified the amino acid frequencies for each of the variable genes within each species to compare residue usage. The mean V-gene, regardless of whether heavy, kappa, or lambda chain, is composed of highly similar quantities of amino acids (Figures 2A, S2A). Furthermore, this amino acid content is highly conserved across the different species analyzed, with the average variable gene often encoding multiple serines ($\geq 12\%$) and low amounts of histidine and methionine ($\sim 0\%$) (Figures 2A, S2A). Surprisingly, the dendrograms of the three loci (heavy, lambda, and kappa) based on hierarchical clustering diverge dramatically from the respective species tree (Figures 2A, 2B, S2A), suggesting selective pressures are more driven towards variable germlines than species.

Given the high similarity of amino acid composition between species, we next compared the distributions of amino acid compositions for the individual variable genes. By performing a t-distributed stochastic neighbor embedding (tSNE) analysis on the amino acid composition matrix and projection onto two dimensions, we were able to visualize the large number of variable genes ($n_{heavy}=718$, $n_{kappa}=454$, $n_{lambda}=254$) (Figure 2C, Figure S4B). If maximal diversity in regards to amino acid usage was selected for, we would expect to observe that the V-genes within a species would be spread out while being in close proximity to V-genes from the other species. Indeed, the majority of the variable heavy segments from seven of the nine species (rabbit and alpaca excluded) did not cluster in a species-specific manner (i.e. nearest neighboring points are often variable segments from different species) (Figure 2B). Interestingly, a species-specific clustering is observed for the kappa variable locus of the rabbit, but not for the lambda segments (Figures S3B), suggesting that the amino acid diversification of variable genes did not undergo identical selective pressures for the various variable segments within the same species.

While the amino acid composition of heavy, kappa and lambda variable segments appeared to be conserved for the nine species analyzed (Figure 3A, 3B, S2A), we questioned if the permutations in which these amino acids occur vary based on species. Therefore, we decomposed the variable segments into k-mers ($k=2,3,4$ strings of amino acids), as this approach has previously been shown to have predictive capabilities of

antibody properties (Greiff et al. 2017b). Interestingly, the variable heavy genes clustered almost exclusively by species after performing tSNE on the matrix of 2-mers, 3-mers, and 4-mers (Figure 2C). Not only was this trend observed for both lambda and kappa variable segments, it actually increased the separation between the two genes (Figure S3C). The two species pairs with the lowest extent of separation were the variable segments for the human and macaque and the mouse and the rat (Figures 2B, 2C), which is not surprising given both species pairs share a relatively recent common ancestor (Figure 2B). Finally, nearly all of the IgH V genes from the two human alleles overlap in the t-SNE projections of both amino acid and k-mer, again suggesting that interspecies diversity is lower than intraspecies diversity. These sequence motifs may influence both the potential for further Ig diversification via somatic hypermutation or gene conversion, and BCR stability, and should be considered when humanizing antibodies discovered in other species.

Species-specific selective pressures of variable germline segments

While phylogenetic trees of the variable Ig segments have been inferred in previous studies (Olivieri et al. 2014), topological metrics specifically describing the selective pressure based on tree shape and branch lengths have, to our knowledge, not yet been applied. Additionally, previous studies have focused on unrooted topologies, which further limits the number of comparative topology metrics (Olivieri et al. 2014). Therefore, we inferred a phylogenetic tree for each variable segment locus (i.e. three trees per species, one for heavy, one for lambda, and one for kappa) using Bayesian phylogenetic inference (BEAST), which produces rooted trees without needing to explicitly specify an ancestral sequence as the outgroup (Drummond and Rambaut 2007; Bouckaert et al. 2014). We first quantified the extent of tree imbalance for the various species, which indicates whether duplication events occurred evenly distributed or if certain segments underwent multiple duplications. To quantify tree imbalance, commonly employed phylogenetic metrics, such as the Colless number, Sackin index, the number of ladders per tree were used (Figure 3A, S3A, S3B). In brief, the Colless number quantifies the imbalance between left and right subtrees at each internal node in a phylogenetic tree, with a lower value indicating a higher degree of imbalance (Colless 1982). The Sackin number describes the number of ancestors for each tip of the tree (Sackin 1972). Lastly, ladders are topological structures that represent a single sequence continuously involved in branching events (Grenfell et al. 2004; Colijn and

Plazzotta 2017). We additionally calculated the maximum distance from the root for each tree, describing the depth of evolution from the inferred ancestral sequence. Overall, the distribution of these metrics varied across heavy, lambda, and kappa variable segments, with some species showing more imbalanced evolution of heavy chains than light chains, and vice versa (Figure 3A, S3A, S3B). Noteworthy, the Colless number, Sackin index, and the mean ladder size were similar for mice and rat heavy chain trees (~ 0.06 , ~ 0.1 , and ~ 0.02 , respectively), and were furthermore lower than in all other species (Figures 3A, 3B, 3C), despite normalizing for the larger number of sequences. Biologically, this implies that the high number of IgH V genes observed in the rat and mouse underwent a more balanced selection, where no single V-gene continuously underwent successive duplication events. The tree imbalance differs dramatically, however, for the rat and mouse lambda chain trees, where mouse phylogenies were perfectly balanced (Colless number=0) and rats have the highest (~ 0.4). This evolutionary divergence is surprising given that both species share a high ratio of IgK V to IgK L expression in vivo (Zhou et al. 1986; Rolink et al. 1991).

In addition to quantifying the selective pressures based on topology shape (Figures 3A, 3B, 3C), we sought to examine the extent of sequence divergence from the most recent common ancestor (MRCA) for each phylogeny. We thereby calculated the maximum distance from the root to each tips (Figure 3D). Without normalizing by the number of sequences, the mouse and the rat showed the highest distance from their respective IgH V ancestral sequence (Figure 3D). This highlights that there is not necessarily a trade-off between phylogenetic breadth (number of unique IgH V genes) and depth (distance from the ancestral sequence) regarding genomic IgH V diversification. We additionally quantified how sensitive these phylogenetic topologies are to allelic diversity by comparing the topological overlap between the IgH V gene phylogenies for both sets of human alleles (Figure 3E). We calculated the Robinson-Foulds distance, a commonly used metric to measure the distance between phylogenetic trees. Despite our previous analyses producing near-identical behavior between the two allelic IgH V gene sets, distinct topologies were observed between the two trees based on visual observation or quantified via the Robinson-Foulds distance (Figure 3E). While few clades were slightly rearranged (e.g. the evolutionary relationship between IgH V3-64 and V8-51), many tips retained their nearest neighboring IgH V sequence across both trees (e.g. IgH V5-78 and 5-51). This inspired us to ask whether other metrics similarly

revealed a distinct phylogenetic fingerprint between species and across the two human IgH V allelic sets.

Laplacian spectra reveal subtle differences in selective pressures between species

In addition to the aforementioned metrics designed to compare topological structure, we applied a recently developed metric capable of comparing unlabeled tree topologies (unlabeled = tree structure ignoring tip labels) based on their underlying Laplacian spectra, which depends upon the structure of the phylogenetic tree (Lewitus and Morlon 2016). This provides a global analysis of the topology while incorporating multiple parameters such as number of species, tree imbalance, and different modalities of subtree structures (Lewitus and Morlon 2016). Visually, these spectral densities varied dramatically between species and loci (Figure 4A). For example, we observed that the human heavy and kappa genes showed different spectral distributions, further suggesting unique selective pressures for each locus (Figure 4A). Noteworthy was the stark contrast between the mouse and rat spectral densities. The mouse heavy chain shows a bimodal distribution (Figure 4A), implying structural inhomogeneity of the underlying tree structure (e.g. some subtrees are much more dense than others) (Lewitus and Morlon 2016). The spectral density arising from the rat variable heavy segments, in contrast, suggests more asymmetrical duplication events (Figure 4A). These Laplacian spectra can then be compared to each other using the Jensen-Shannon distance, which quantifies divergences between probability densities (Wong and You 1985). Although there were visual differences between the spectral densities of the mouse and rat heavy segment phylogenies (Figure 4A), the heavy chains of the two species clustered next to each other based on the Jansen-Shannon distance (Figure 4A). Similarly, the two human allele sets and the macaque IgH V genes clustered closely together, as expected from visual observation of the Laplacian spectra (Figure 4A). Surprisingly, the zebrafish IgH V genes clustered closer to the two sets of human IgH V genes than the macaque, likely due to the lack of a bimodal seen in the macaque's Laplacian spectra (Figure 4A). While neither a species-specific or chain-specific clustering was observed (Figure S4A), this analysis revealed detectable differences between even closely related sets of IgH V genes. We finally asked if there were closely related sequences shared between these species that had similar Laplacian spectra by inferring a cross-species tree of all human, macaque, platypus, and zebrafish IgH V genes (Figure S5). While there were many germline segments conserved between

human and the macaque, there was almost no overlap with the sequences from either the platypus or zebrafish (Figure S5). This suggests that the similar Laplacian spectra between the two species was entirely due to the pattern of selection regarding which genes underwent duplication events instead of the conservation of highly similar genes. This lack of conservation revealed by the cross-donor phylogenies inspired us to further characterize the genetic distance between the various species.

High amino acid sequence diversity for variable heavy genes

While comparing phylogenetic trees provides insight into the selective pressures faced during segment duplication events, the extent of sequence diversification within a given tree is not immediately clear. Therefore, we utilized the edit distance to quantify the number of amino acid mutations required to convert one germline sequence into the other (Levenshtein I 1966). We hypothesized that all species would show a broad distribution of edit distances between the variable segments to maximize the sequence space. As an initial example to visualize this distribution of diversity, we calculated the pairwise edit distance between all human IgH V segments for both sets of included IMGT alleles (*01 and *02). For example, the minimum distance separating two distinct (i.e. two non-allelic IMGT entries both with *01) V genes (IgH V3–53 and V3–66) was only one amino acid mutation, whereas the maximum distance was ~70 amino acids (IgH V2–26 and V1–68). We subsequently performed a similar analysis for the variable segments (heavy, kappa, and lambda) of the remaining sequences (Figures 5B, S4). While the distribution of mutations separating the heavy chains of most species (human, macaque rabbit, mouse, platypus, rat, and zebrafish) show a continuous density distribution, two species (dog and alpaca) show multimodal distributions of edit distances (Figure 5B). Surprisingly, the mouse and rat lambda genes had the highest median pairwise edit distance (Figure S3, median = ~75 amino acid mutations), despite their low use compared to kappa light chains *in vivo* (Zhou et al. 1986). This suggests that the diversity is maximized for those rare events when the lambda locus is selected during somatic recombination.

We next questioned whether this intraspecies diversity was smaller than the interspecies diversity. We thereby calculated the pairwise amino acid edit distance between all IgH V genes between the two human allelic sets, human and macaque and human and zebrafish, respectively (Figure 5C). We observed similar distributions of edit distances for the human allelic sets, and the human to macaque comparison, whereas the median

edit distance separating IgH V genes from humans and zebrafish was much higher. We subsequently calculated the median distance between the variable heavy chains for each permutation of two species, in addition to the median distance between germlines within a single species (Figure 5D). For all species excluding the rat, the median intraspecies (displayed by the diagonal of the heatmap) edit distance between heavy segments was smaller than the interspecies edit distance (Figure 5D), which is consistent with the clustering we found based on k-mers (Figure 2C, Figure S2C). Those species with the lowest amount of intraspecies diversity (platypus, alpaca, and dog) also exhibited a low degree of interspecies diversity (Figure 5C). We finally sought to quantify the edit distance (both nucleotide and amino acid) of those IgH V alleles corresponding to the same V gene in the two human germline sets. As previously mentioned, 13 of the analyzed IgH V genes here were common to both human_01 and human_02 due to lack of IgH V genes with annotated alleles. Although the remaining 50 alleles possessed unique nucleotide sequences between the two human germline sets, there were 12 additional allelic variants that had identical amino acid sequences (Figure 5D) despite different nucleotide sequences. Furthermore, the median difference between corresponding IgH V genes between human allele set 01 and 02 was only one amino acid mutation and two nucleotide mutations (after excluding the 13 shared IgH V genes).

Antigen-independent germline diversity varies across different murine B cell populations

All of our previous analyses were based on germline diversity relationships independent of B cells and recombination, therefore, we lastly addressed whether germline-encoded diversity signatures could be detected in various B cell populations after recombination and selective pressures (e.g., antigen challenge). We utilized the sequencing data from more than 20 previously conducted Ig-Seq experiments examining the heavy chain repertoires of naïve B cells, pre-B cells, and plasma cells under four different antigen conditions (no antigen (n=5), ovalbumin (OVA, n=5), 4-hydroxy-3-nitrophenylacetyl conjugated to hen egg lysozyme (NP-HEL, n=5, and hepatitis B virus surface antigen (HBsAg, n=4)) (Greiff et al. 2017a). We then calculated the pairwise amino acid edit distances between the top 20 heavy variable genes for each cohort (Figure 6). We observed the broadest distributions for the pre-B cell and the naïve B cell compartments, which are prior to antigen challenge and

selection (Figure 6). Interestingly, those mice stimulated with OVA showed an increase in pairwise distances of both the pre-B cell ($p=2.4 \times 10^{-6}$) and naive B cell ($p=4.3 \times 10^{-5}$) compartments compared to the other three antigen cohorts, suggesting the potential influence of antigen on germline gene use. For the plasma cell cohorts, IgG and IgM isotypes were analyzed separately despite coming from the same experimental sample (Greiff et al. 2017a). The pairwise edit distances observed in bone marrow were significantly increased for the IgM isotype compared to the IgG isotype for all four antigen cohorts ($p=7.9 \times 10^{-5}$, $< 2.2 \times 10^{-16}$, $< 2.2 \times 10^{-16}$, $< 2.2 \times 10^{-16}$). Surprisingly, the difference between the IgG and IgM compartment for the splenic plasma cells was much less profound compared to the bone marrow plasma cells (excluding the HBsAg cohort). Furthermore, the top twenty V genes showed significantly less germline-encoded diversity for both IgG plasma cell compartments (bone marrow and spleen) in the HBsAg cohort compared to the other antigens (Figure 6). Overall, these findings suggest that genomic diversity arising from germline elements is reflected in the expressed repertoire and is variable based on B cell population.

Discussion

Leveraging the increased amount of publicly available genomic and Ig-Seq data, we have provided a comparative analysis of the germline-encoded diversity of the immunoglobulin variable, diversity, and joining segments across multiple species. Specifically, we have shown that sequence diversification and germline gene segment duplication events have largely occurred in a species-specific manner that diverges from the underlying species tree of life (Figure 2B). Understanding the rules of immunoglobulin selection and the resulting sequence diversity can improve our understanding of how the humoral immune response successfully interacts with a nearly limitless number of potential foreign antigens.

We have additionally provided phylogenetic analyses describing the selective pressures of the variable heavy, lambda, and kappa chains for each species. We could thereby quantify how often various variable gene segments underwent duplication events based on the inferred topology. The V-genes that have undergone multiple duplications may be better candidates for antibody design given their proven robustness to naturally occurring mutations. Furthermore, metrics such as the Laplacian spectra revealed subtle differences between the selective pressures experienced by the rat and the mouse, which are often assumed to share a high degree of homology. While we have performed this analysis on the entire variable region of the variable germline segments, it is possible that separating variable genes into their complementary determining regions (CDR) and framework regions (FR) could highlight even more species-specific patterns of selection. Indeed, previous studies have noted differences in both diversity and selection between the two regions for humans (Romo-González and Vargas-Madrado 2005; Schwartz and Hershberg 2013), warranting further analysis for other species.

Next, we examined how variable gene diversity is distributed, both in an intra- and interspecies manner (Figure 5). For example, some of the human IgH V genes showed extreme degrees of sequence similarity despite mapping to unique loci in the genome (Figure 5A). These variable genes can be separated by as little as a single amino acid, which is also the case for some allelic variants across individuals within the same species (Figure 5E), which has important ramifications for aligning sequencing reads from Ig-seq data to reference germline databases for distinct biological samples. Furthermore, we observed comparable sequence similarity across two distinct allelic sets of human IgH V genes (Figures 5A, 5B), suggesting that sequence diversity between IgH V genes is maintained within an individual. Multiple metrics supported

424 this breadth of diversification including GC content, protein solubility, and protein
425 stability (Figures 1C, 1E, 1F). Our analysis revealed minor differences between two
426 sets of human IgH V alleles, as indicated by highly correlated biophysical properties
427 (Figure 1), closely-clustered IgH V genes following both amino acid and k-mer analysis
428 (Figure 2), convergent phylogenetic fingerprints (Figures 3, 4), and overlapping edit
429 distance distributions (Figure 5). A comparably high degree of similarity between the
430 two human allelic IgH V sets was also observed when comparing the IgH V genes
431 between humans and macaques. Likewise, mouse and rat V genes clustered similarly
432 across multiple metrics compared to that with other species, again, unsurprising given
433 their relatively minor evolutionary distance.

434 Finally, we applied sequence diversity metrics to Ig-Seq data that characterized
435 different murine B cell populations across different immunization cohorts. While V-
436 gene frequencies and somatic hypermutation has been analyzed extensively across
437 many different experimental settings, little effort has been made to quantify the
438 germline-encoded diversity in these Ig-Seq experiments. Our analysis revealed that the
439 top V-genes used in the pre-B cell and naïve B cell populations showed a robust
440 increase of germline-encoded diversity when comparing to class-switched IgG plasma
441 cells (Figure 6). In contrast to these B cell populations, the specific antigen (with
442 exception of the HBsAg immunized cohort) had no discernable influence on the V gene
443 diversity within IgG plasma cells from spleen and bone marrow, suggesting structured
444 diversity across different B cell populations via germline gene usage.

445 One major limitation to this and other comparative studies involving adaptive immune
446 germline genes is that the annotations for many species may be incomplete and that the
447 sampling of germline genes in the different species varies. While we have normalized
448 by the number of genes being analyzed in each analysis when possible, considerable
449 effects arising due to variation regarding how well a given species' germline is
450 annotated. Our study is further influenced by differences in variable germline segments
451 between strains of the same species (Collins et al. 2015), as it has already been shown
452 that the IgH germline alone can influence antigen-specific response and memory
453 formation to phycoerythrin (Pape et al. 2018). Additionally, we have attempted to
454 restrict our analysis to a single set of alleles per locus (i.e. always the *01 allele from
455 IMGT or from a single genome for the variable light chain segments), thereby ignoring
456 the existence of haplotypes within each species, despite recent evidence implicating
457 their presence (Kidd et al. 2012; Corcoran et al. 2016; Kirik et al. 2017; Watson et al.

2017). The haplotypic diversity will continue to be better characterized as the availability of personalized genome sequencing expands. While we have included two sets of human alleles for the IgH V genes to estimate the sequence diversity within a given species, this is but a crude estimation in reality, again due to the difficulty in annotating variable human germline genes. Indeed, genetic variation between individuals of the same species has been well-established (Sasso et al. 1995; Ching et al. 2005; Boyd et al. 2010; Watson et al. 2015, 2017). Additionally, the reference alleles are not necessarily arising from a single individual, even if the majority of the alleles can be traced back to a single accession number. Finally, we acknowledge that we have not verified the accuracy of all human alleles in this study and admit that there may be sequence or annotation errors in currently utilized databases. Such erroneous annotations may be extended to pseudogenes for many species, as this information is difficult to infer from those species less commonly studied. These crucial limitations, however, highlight the necessity to better characterization of the entire immunoglobulin variable region within each individual, as opposed to databases mixing variable gene segments from multiple individuals. Continued work by the adaptive immune receptor repertoire (AIRR) community to systematically store and annotate the germline sequences for certain species represents crucial progress to better comprehend the genomic selection underlying germline diversity (Breden et al. 2017; Ohlin et al. 2019; Lees et al. 2019; Omer et al. 2019). Additionally, Advancements of highly accurate and long-read sequencing technologies are already being incorporated to better understand how germline diversity differs within the population (Watson et al. 2013). Indeed, others have already displayed that specific germlines preferentially interact with particular antigens (Avnir et al. 2016; Jardine et al. 2016; Pieper et al. 2017), suggesting that further characterizing genomic diversity can be used for structure-guided vaccination and antibody discovery.

Another potential bias in our analysis is that we obtained the reference IgH and IgK/L germline segments from different sources due to multiple included species lacking light chain information on IMGT. Thereby we acquired the light chain amino acid sequences from Vgenerepertoire (Olivieri et al. 2014), which infers variable segments from whole genome sequencing. To minimize the influence of this discrepancy, we have primarily focused upon the heavy chain variable region due to its importance in dictating clonality and binding-specificity (Xu and Davis 2000; Zhou and Kleinstein 2019). Future work,

491 however, could quantify if and how erroneous *de novo* annotation of variable germline
492 segments is when using newer whole genome sequencing technologies.

493 Since our study shows the value for exhaustively characterizing germline-encoded
494 diversity, we have developed an R package, VDJgermlines, which both stores BCR and
495 TCR alleles for a variety of species and enables the retrieval of variable germlines (both
496 TCR and BCR) based on species, chain, database, and strain. We hold that this will
497 help future research explore Ig germline diversity and selection and elucidate the
498 ramifications genomic diversity has on repertoire diversity. Furthermore, many
499 different species have been employed for the discovery and engineering of antibodies
500 for biotechnology applications (e.g., reagents, diagnostics, therapeutics), highlighting
501 the importance of characterizing how antibody repertoire diversity is generated via
502 somatic diversity. The necessity of humanizing monoclonal antibodies for therapeutic
503 purposes clearly displays the importance that species-specific diversity can have on
504 determining the success of monoclonal antibody treatments. We have shown that
505 patterns of amino acids can be used to separate the majority of V-genes by species,
506 despite each species having similar frequencies of the underlying amino acid
507 distribution. This information may prove relevant to improve the humanization of
508 antibodies by incorporating as many human-like motifs as possible without disrupting
509 antigen specificity.

510

Materials and Methods

Germline genes

The heavy chain variable, diversity and joining segments for all species were taken from the IMGT reference directory (Lefranc et al. 1999) (October 2, 2019). Only IgH V genes with IMGT-annotations “C57B6” and “Sprague-Dawley” were included in the mouse and rat analysis, respectively. Those IgH genes annotated as variants with open reading frames on IMGT (“OR” in gene name in IMGT reference directory) were removed. For all species other than human, only the first allele for each unique IgH V gene (“*01”) was selected. For the germline set annotated as human_01, only human IgH genes annotated with *01 on IMGT were selected. For the germline set annotated with human_02, all human IgH genes with *02 in their IMGT annotation were selected. The same IgH V gene was used for both reference set (n=13) if it was only found within one set of human IgH V alleles. Any remaining sequences labeled as pseudogenes (“P”) in the IMGT annotation after filtering for frameshifts were included in the main analysis due to some species having a majority or all of annotated IgH V genes labeled as pseudogenes. The kappa and lambda chains for the nine species were downloaded from a single accession number from Vgenerepertoire (Olivieri et al. 2014) due to multiple species entirely lacking light chain gene segments on IMGT (e.g. platypus, alpaca, crab-eating macaque). The name and accession number of the each gene analyzed in this study is included in supplementary tables 1 and 2. Additionally, full-length nucleotide and amino sequences, in addition to V gene names and accession number, for all species are can be easily accessed in our R package VDJgermlines, which is publicly available and stored on CRAN.

Inferred biophysical properties

The overall solubility score for each included IgH V gene was calculated using the online CamSol server with the pH set to 7 (Sormanni et al. 2015). Protein stability was inferred using the functions *aIndex* and *instaindex* in the R Peptides package (Osorio, D., Rondon-Villarreal, P. & Torres, R 2015) to calculate the Ikai aliphatic (1980) and instability indices (Guruprasad et al. 1990), respectively.

k-mer and tSNE analysis

The variable genes for all species were decomposed into their constituent overlapping kmers (k=2,3, and 4 patterns of amino acids) using the KeBABS package (Palme et al. 2015). The subsequent matrices were appended to each other (column-wise) and were used as input using the R package Rtsne using with the perplexity parameter set to 20 (version 0.13).

Phylogenetic analysis

BEAST (version 2.5) was used to infer phylogenetic trees separately for the variable heavy, lambda, and kappa chains for each species. The Dayhoff substitution model was used with gamma-distributed site heterogeneity. A relaxed lognormal clock model was used to estimate divergence times, with the number of discrete rates set to -1 and the clock rate parameter set to 1. The Yule model prior was used with a uniformly distributed birth rate and a gamma distributed uclDStdev. The chain length of the Markov chain Monte Carlo algorithm was set to 100,000,000 iterations to ensure convergence of the posterior probability distributions. The maximum clade credibility (MCC) tree was extracted using the program TreeAnnotator (Bouckaert et al. 2014), with a 10% burn-in to ensure that the analyzed trees had reached equilibrium in the posterior distribution.

The average number of ladders per tree, the Colless number, and the Sackin index were calculated using the R package PhyloTop (version 2.0.1). The Laplacian spectra were calculated using the R package RPANDA (version 1.3) using the spectR function with each phylogenetic tree as input. The spectra were plotted using the function plot_spectR. The Jensen Shannon index was calculated using the function JSDtree with the default parameters. The species-tree was inferred using phyloT and subsequently plotted via the Interactive Tree of Life software (iTOL) (Letunic and Bork 2016).

Ig-Seq analysis

Pre-processing of the antibody repertoire sequencing data was performed as previously reported (Greiff et al. 2017a). Briefly, sequence reads from an immunization previously performed immunization study involving four different murine cohorts (Ovalbumin (OVA), 4-hydroxy-3-nitrophenylacetyl conjugated to hen egg lysozyme (NP), no immunization (Control) or hepatitis B virus surface antigen (HBsAg) were aligned and processed using VDJ alignment clonotyping using the MiXCR software package (clonotyping based on the CDRH3 region) (Bolotin et al.

2015). For plasma cell samples (bone marrow and spleen), reads were separated based on inferred isotype (IgG or IgM) and analyzed separately. The average V gene frequencies based on unique clones was calculated for each immunization cohort (4-5 mice per vaccine cohort). The top 20 V genes (regardless of whether they were the same V genes between cohorts) were selected and the pairwise Levenshtein distances between the germline-encoded segments (i.e. ignoring any CDRH3 rearrangements or somatic hypermutations that may have occurred) were calculated using the *stringdistmatrix* function in the R package stringdist (Loo 2014).

Determination of statistical significance

Significance between groups was tested using the Wilcoxon rank-sum test.

References

- Abbott RK, Lee JH, Menis S, et al (2018) Precursor Frequency and Affinity Determine B Cell Competitive Fitness in Germinal Centers, Tested with Germline-Targeting HIV Vaccine Immunogens. *Immunity* 48:133-146.e6. <https://doi.org/10.1016/j.immuni.2017.11.023>
- Avnir Y, Watson CT, Glanville J, et al (2016) IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Scientific Reports* 6:20842. <https://doi.org/10.1038/srep20842>
- Bolotin DA, Poslavsky S, Mitrophanov I, et al (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat Meth* 12:380–381. <https://doi.org/10.1038/nmeth.3364>
- Bouckaert R, Heled J, Kühnert D, et al (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology* 10:e1003537. <https://doi.org/10.1371/journal.pcbi.1003537>
- Boyd SD, Gaëta BA, Jackson KJ, et al (2010) Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements. *Jl* 184:6986–6992. <https://doi.org/10.4049/jimmunol.1000445>
- Breden F, Luning Prak ET, Peters B, et al (2017) Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front Immunol* 8:1418. <https://doi.org/10.3389/fimmu.2017.01418>
- Brown AJ, Snapkov I, Akbar R, et al (2019) Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of

614 adaptive immune receptor repertoires. *Mol Syst Des Eng* 4:701–736.
 615 <https://doi.org/10.1039/C9ME00071B>

616 Chimge N-O, Pramanik S, Hu G, et al (2005) Determination of gene organization
 617 in the human IGHV region on single chromosomes. *Genes Immun* 6:186–
 618 193. <https://doi.org/10.1038/sj.gene.6364176>

619 Colijn C, Plazzotta G (2017) A Metric on Phylogenetic Tree Shapes. *Systematic*
 620 *Biology*. <https://doi.org/10.1093/sysbio/syx046>

621 Colless DH (1982). *Systematic Zoology* 31:100–104.
 622 <https://doi.org/10.2307/2413420>

623 Collins AM, Wang Y, Roskin KM, et al (2015) The mouse antibody heavy chain
 624 repertoire is germline-focused and highly variable between inbred
 625 strains. *Phil Trans R Soc B* 370:20140236.
 626 <https://doi.org/10.1098/rstb.2014.0236>

627 Corcoran MM, Phad GE, Bernat NV, et al (2016) Production of individualized V
 628 gene databases reveals high levels of immunoglobulin genetic diversity.
 629 *Nature Communications* 7:13642.
 630 <https://doi.org/10.1038/ncomms13642>

631 Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by
 632 sampling trees. *BMC Evolutionary Biology* 7:214.
 633 <https://doi.org/10.1186/1471-2148-7-214>

634 Elhanati Y, Murugan A, Callan CG, et al (2014) Quantifying selection in immune
 635 receptor repertoires. *PNAS* 111:9875–9880

636 Elhanati Y, Sethna Z, Marcou Q, et al (2015) Inferring processes underlying B-cell
 637 repertoire diversity. *Philosophical Transactions of the Royal Society B:*
 638 *Biological Sciences* 370:20140243.
 639 <https://doi.org/10.1098/rstb.2014.0243>

640 Friedensohn S, Lindner JM, Cornacchione V, et al (2018) Synthetic Standards
 641 Combined With Error and Bias Correction Improve the Accuracy and
 642 Quantitative Resolution of Antibody Repertoire Sequencing in Human
 643 Naïve and Memory B Cells. *Front Immunol* 9:.
 644 <https://doi.org/10.3389/fimmu.2018.01401>

645 Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH (2015) Automated analysis of
 646 high-throughput B-cell sequencing data reveals a high frequency of novel
 647 immunoglobulin V gene segment alleles. *PNAS* 112:E862--E870.
 648 <https://doi.org/10.1073/pnas.1417683112>

649 Georgiou G, Ippolito GC, Beausang J, et al (2014) The promise and challenge of
 650 high-throughput sequencing of the antibody repertoire. *Nat Biotech*
 651 32:158–168. <https://doi.org/10.1038/nbt.2782>

652 Gidoni M, Snir O, Peres A, et al (2019) Mosaic deletion patterns of the human
 653 antibody heavy chain gene locus shown by Bayesian haplotyping. Nat
 654 Commun 10:628. <https://doi.org/10.1038/s41467-019-08489-3>

655 Glanville J, Zhai W, Berka J, et al (2009) Precise determination of the diversity of
 656 a combinatorial antibody library gives insight into the human
 657 immunoglobulin repertoire. Proceedings of the National Academy of
 658 Sciences 106:20216–20221. <https://doi.org/10.1073/pnas.0909775106>

659 Greiff V, Bhat, P, Cook SC, et al Accurate detection of immunological status by
 660 immune repertoire diversity profiling. Submitted

661 Greiff V, Menzel U, Miho E, et al (2017a) Systems analysis reveals high genetic
 662 and antigen-driven predetermination of antibody repertoires throughout
 663 B-cell development. Cell Reports

664 Greiff V, Miho E, Menzel U, Reddy ST (2015) Bioinformatic and Statistical
 665 Analysis of Adaptive Immune Repertoires. Trends in Immunology
 666 36:738–749. <https://doi.org/10.1016/j.it.2015.09.006>

667 Greiff V, Weber CR, Palme J, et al (2017b) Learning the High-Dimensional
 668 Immunogenomic Features That Predict Public and Private Antibody
 669 Repertoires. The Journal of Immunology 199:2985–2997.
 670 <https://doi.org/10.4049/jimmunol.1700594>

671 Grenfell BT, Pybus OG, Gog JR, et al (2004) Unifying the Epidemiological and
 672 Evolutionary Dynamics of Pathogens. Science 303:327.
 673 <https://doi.org/10.1126/science.1090727>

674 Guruprasad K, Reddy BVB, Pandit MW (1990) Correlation between stability of a
 675 protein and its dipeptide composition: a novel approach for predicting *in*
 676 *vivo* stability of a protein from its primary sequence. Protein Eng Des Sel
 677 4:155–161. <https://doi.org/10.1093/protein/4.2.155>

678 Howe K, Clark MD, Torroja CF, et al (2013) The zebrafish reference genome
 679 sequence and its relationship to the human genome. Nature 496:498–503.
 680 <https://doi.org/10.1038/nature12111>

681 Janeway CA, Murphy K (2011) Janeway's Immunobiology, 8th Revised edition.
 682 Taylor & Francis

683 Jardine JG, Kulp DW, Havenar-Daughton C, et al (2016) HIV-1 broadly
 684 neutralizing antibody precursor B cells revealed by germline-targeting
 685 immunogen. Science 351:1458–1463.
 686 <https://doi.org/10.1126/science.aad9195>

687 Johnston CM, Wood AL, Bolland DJ, Corcoran AE (2006) Complete Sequence
 688 Assembly and Characterization of the C57BL/6 Mouse Ig Heavy Chain V
 689 Region. J Immunol 176:4221–4234.
 690 <https://doi.org/10.4049/jimmunol.176.7.4221>

691 Kidd MJ, Chen Z, Wang Y, et al (2012) The inference of phased haplotypes for the
692 immunoglobulin H chain V region gene loci by analysis of VDJ gene
693 rearrangements. *J Immunol* 188:1333–1340.
694 <https://doi.org/10.4049/jimmunol.1102097>

695 Kirik U, Greiff L, Levander F, Ohlin M (2017) Parallel antibody germline gene and
696 haplotype analyses support the validity of immunoglobulin germline gene
697 inference and discovery. *Molecular Immunology* 87:12–22.
698 <https://doi.org/10.1016/j.molimm.2017.03.012>

699 Kudla G, Lipinski L, Caffin F, et al (2006) High Guanine and Cytosine Content
700 Increases mRNA Levels in Mammalian Cells. *PLoS Biology* 4:e180.
701 <https://doi.org/10.1371/journal.pbio.0040180>

702 Laserson U, Vigneault F, Gadala-Maria D, et al (2014) High-resolution antibody
703 dynamics of vaccine-induced immune responses. *PNAS* 111:4928–4933.
704 <https://doi.org/10.1073/pnas.1323862111>

705 Lavinder JJ, Hoi KH, Reddy ST, et al (2014) Systematic Characterization and
706 Comparative Analysis of the Rabbit Immunoglobulin Repertoire. *PLoS*
707 *ONE* 9:e101322. <https://doi.org/10.1371/journal.pone.0101322>

708 Lees W, Busse CE, Corcoran M, et al (2019) OGRDB: a reference database of
709 inferred immune receptor genes. *Nucleic Acids Research* gkz822.
710 <https://doi.org/10.1093/nar/gkz822>

711 Lefranc M-P, Giudicelli V, Ginestoux C, et al (1999) IMGT, the international
712 ImMunoGeneTics database. *Nucl Acids Res* 27:209–212.
713 <https://doi.org/10.1093/nar/27.1.209>

714 Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the
715 display and annotation of phylogenetic and other trees. *Nucleic Acids*
716 *Research* 44:W242–W245. <https://doi.org/10.1093/nar/gkw290>

717 Levenshtein I V (1966) Binary codes capable of correcting deletions, insertions,
718 and reversals. *Soviet Physics* 10:707–710

719 Lewitus E, Morlon H (2016) Characterizing and Comparing Phylogenies from
720 their Laplacian Spectrum. *Systematic Biology* 65:495–507.
721 <https://doi.org/10.1093/sysbio/syv116>

722 Loo MPJ van der (2014) The stringdist package for approximate string matching.
723 *The R Journal* 6:111–122

724 Methot SP, Di Noia JM (2017) Molecular Mechanisms of Somatic Hypermutation
725 and Class Switch Recombination. In: *Advances in Immunology*. Elsevier,
726 pp 37–87

727 Miho E, Yermanos A, Weber CR, et al (2018) Computational Strategies for
728 Dissecting the High-Dimensional Complexity of Adaptive Immune

729 Repertoires. *Frontiers in Immunology* 9:.
730 <https://doi.org/10.3389/fimmu.2018.00224>

731 Nadel B, Feeney AJ (1997) Nucleotide deletion and P addition in V (D) J
732 recombination: a determinant role of the coding-end sequence. *Molecular*
733 and cellular biology 17:3768–3778

734 Ohlin M, Scheepers C, Corcoran M, et al (2019) Inferred Allelic Variants of
735 Immunoglobulin Receptor Genes: A System for Their Evaluation,
736 Documentation, and Naming. *Front Immunol* 10:435.
737 <https://doi.org/10.3389/fimmu.2019.00435>

738 Olivieri DN, von Haeften B, Sánchez-Espinel C, et al (2014) Genomic V exons from
739 whole genome shotgun data in reptiles. *Immunogenetics* 66:479–492.
740 <https://doi.org/10.1007/s00251-014-0784-3>

741 Omer A, Shemesh O, Peres A, et al (2019) VDJbase: an adaptive immune receptor
742 genotype and haplotype database. *Nucleic Acids Research* gkz872.
743 <https://doi.org/10.1093/nar/gkz872>

744 Osorio, D., Rondon-Villarreal, P. & Torres, R (2015) Peptides: A package for data
745 mining of antimicrobial peptides. *The R Journal* 7:4–14

746 Palme J, Hochreiter S, Bodenhofer U (2015) KeBABS: an R package for kernel-
747 based analysis of biological sequences. *Bioinformatics* btv176.
748 <https://doi.org/10.1093/bioinformatics/btv176>

749 Pape KA, Maul RW, Dileepan T, et al (2018) Naive B Cells with High-Avidity
750 Germline-Encoded Antigen Receptors Produce Persistent IgM + and
751 Transient IgG + Memory B Cells. *Immunity*.
752 <https://doi.org/10.1016/j.immuni.2018.04.019>

753 Parola C, Neumeier D, Reddy ST (2018) Integrating high-throughput screening
754 and sequencing for monoclonal antibody discovery and engineering.
755 *Immunology* 153:31–41. <https://doi.org/10.1111/imm.12838>

756 Pieper K, Tan J, Piccoli L, et al (2017) Public antibodies to malaria antigens
757 generated by two LAIR1 insertion modalities. *Nature* 548:597–601.
758 <https://doi.org/10.1038/nature23670>

759 Ralph DK, Iv FAM (2016) Likelihood-Based Inference of B Cell Clonal Families.
760 *PLOS Computational Biology* 12:e1005086.
761 <https://doi.org/10.1371/journal.pcbi.1005086>

762 Reddy ST, Ge X, Miklos AE, et al (2010) Monoclonal antibodies isolated without
763 screening by analyzing the variable-gene repertoire of plasma cells. *Nat*
764 *Biotech* 28:965–969. <https://doi.org/10.1038/nbt.1673>

765 Rolink A, Streb M, Melchers F (1991) The λ/λ ratio in surface immunoglobulin
766 molecules on B lymphocytes differentiating from DHJH-rearranged

767 murine pre-B cell clones in vitro. *European Journal of Immunology*
768 21:2895–2898. <https://doi.org/10.1002/eji.1830211137>

769 Romo-González T, Vargas-Madrado E (2005) Structural analysis of substitution
770 patterns in alleles of human immunoglobulin VH genes. *Molecular*
771 *Immunology* 42:1085–1097.
772 <https://doi.org/10.1016/j.molimm.2004.11.004>

773 Sackin MJ (1972) “Good” and “Bad” Phenograms. *Systematic Zoology* 21:225–
774 226. <https://doi.org/10.2307/2412292>

775 Sasso EH, Buckner JH, Suzuki LA (1995) Ethnic differences of polymorphism of
776 an immunoglobulin VH3 gene. *J Clin Invest* 96:1591–1600

777 Schiaffella E, Sehgal D, Anderson AO, Mage RG (1999) Gene Conversion and
778 Hypermutation During Diversification of V_H Sequences in Developing
779 Splenic Germinal Centers of Immunized Rabbits. *J Immunol* 162:3984

780 Schwartz GW, Hershberg U (2013) Conserved variation: identifying patterns of
781 stability and variability in BCR and TCR V genes with different diversity
782 and richness metrics. *Phys Biol* 10:035005.
783 <https://doi.org/10.1088/1478-3975/10/3/035005>

784 Sormanni P, Amery L, Ekizoglou S, et al (2017) Rapid and accurate in silico
785 solubility screening of a monoclonal antibody library. *Sci Rep* 7:8200.
786 <https://doi.org/10.1038/s41598-017-07800-w>

787 Sormanni P, Aprile FA, Vendruscolo M (2015) The CamSol Method of Rational
788 Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular*
789 *Biology* 427:478–490. <https://doi.org/10.1016/j.jmb.2014.09.026>

790 Stanfield RL, Wilson IA, Smider VV (2016) Conservation and diversity in the
791 ultralong third heavy-chain complementarity-determining region of
792 bovine antibodies. *Science Immunology* 1:aaf7962–aaf7962.
793 <https://doi.org/10.1126/sciimmunol.aaf7962>

794 Tonegawa S (1983) Somatic generation of antibody diversity. *Nature* 302:575–
795 581. <https://doi.org/10.1038/302575a0>

796 Vollmers C, Sit RV, Weinstein JA, et al (2013) Genetic measurement of memory B-
797 cell recall using antibody repertoire sequencing. *Proceedings of the*
798 *National Academy of Sciences* 110:13463–13468.
799 <https://doi.org/10.1073/pnas.1312146110>

800 Watson CT, Glanville J, Marasco WA (2017) The Individual and Population
801 Genetics of Antibody Immunity. *Trends in Immunology* 38:459–470.
802 <https://doi.org/10.1016/j.it.2017.04.003>

803 Watson CT, Steinberg KM, Graves TA, et al (2015) Sequencing of the human IG
804 light chain loci from a hydatidiform mole BAC library reveals locus-

805 specific signatures of genetic diversity. *Genes Immun* 16:24–34.
 806 <https://doi.org/10.1038/gene.2014.56>

807 Watson CT, Steinberg KM, Huddleston J, et al (2013) Complete Haplotype
 808 Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity,
 809 and Joining Genes and Characterization of Allelic and Copy-Number
 810 Variation. *Am J Hum Genet* 92:530–546.
 811 <https://doi.org/10.1016/j.ajhg.2013.03.004>

812 Weber J, Peng H, Rader C (2017) From rabbit antibody repertoires to rabbit
 813 monoclonal antibodies. *Exp Mol Med* 49:e305.
 814 <https://doi.org/10.1038/emm.2017.23>

815 Weinstein JA, Jiang N, White RA, et al (2009) High-Throughput Sequencing of the
 816 Zebrafish Antibody Repertoire. *Science* 324:807–810.
 817 <https://doi.org/10.1126/science.1170020>

818 Wong A, You M (1985) Entropy and Distance of Random Graphs with Application
 819 to Structural Pattern Recognition. *EEE Transactions on Pattern Analysis*
 820 and Machine Intelligence PAMI-7 599–609

821 Xu JL, Davis MM (2000) Diversity in the CDR3 Region of VH Is Sufficient for Most
 822 Antibody Specificities. *Immunity* 13:37–45.
 823 [https://doi.org/10.1016/S1074-7613\(00\)00006-6](https://doi.org/10.1016/S1074-7613(00)00006-6)

824 Yermanos A, Greiff V, Krautler NJ, et al (2017) Comparison of methods for
 825 phylogenetic B-cell lineage inference using time-resolved antibody
 826 repertoire simulations (AbSim). *Bioinformatics* 33:3938–3946.
 827 <https://doi.org/10.1093/bioinformatics/btx533>

828 Zhou D, Leslie GA, Guo K, Gutman GA (1986) EXPRESSION OF
 829 IMMUNOGLOBULIN LAMBDA CHAINS IN THE LABORATORY RAT.
 830 *European Journal of Immunogenetics* 13:299–308.
 831 <https://doi.org/10.1111/j.1744-313X.1986.tb01114.x>

832 Zhou JQ, Kleinstein SH (2019) Cutting Edge: Ig H Chains Are Sufficient to
 833 Determine Most B Cell Clonal Relationships. *Ji* 203:1687–1692.
 834 <https://doi.org/10.4049/jimmunol.1900666>

835 (1980) Thermostability and Aliphatic Index of Globular Proteins. *The Journal of*
 836 *Biochemistry*. <https://doi.org/10.1093/oxfordjournals.jbchem.a133168>

837
 838

Figure 1. Different number and biochemical properties of IgH V genes across nine species. (A) The number of in-frame IgH V genes included in this study for each of nine commonly-studied species. Samples labeled “Human_01” and “Human_02” include to all productive IgH V genes from the human IMGT reference sequence containing allele 01 or 02, respectively. (B) The overlap of IgH V nucleotide sequences shared across both included human allele sets. For 13 IgH V genes unique to one of the allelic sets, the corresponding sequence from the other allelic set was selected. (C) Violin plots of the GC content across all IgH V genes for each species. Dashed lines indicate median and solid lines indicate quartiles. (D) Correlation between GC content from allelic IgH V sets Human_01 and Human_02. Filled squares (n=50) indicate unique sequences and open triangles (n=13) indicate identical sequences between samples. (E) Violin plots of solubility score for all IgH V genes. (F) Violin plots of aliphatic index for all IgH V genes. (G) Correlation between CamSol solubility score from allelic IgH V sets. (H) Correlation between aliphatic index from allelic IgH V sets.

Figure 2. Amino acid motifs separate variable genes by species. (A) Mean amino acid frequencies for variable heavy germline segments. The intensity corresponds to the mean occurrence (%) of a given amino acid in the IgH V gene segments for each species. (B) Species-tree of the nine species employed in this study compared to the dendrogram. (C) t-SNE plot projecting the amino acid frequencies for the individual variable heavy segments into two dimensions. Each dot represents a single variable heavy chain. (D) t-SNE plot projecting the k-mers (k=2,3,4) composition of each variable chain into two dimensions. Each dot represents a single variable heavy chain.

Figure 3. Evolutionary topologies reveal species-specific section for variable heavy immunoglobulin germline genes. (A-D) The normalized Colless number, normalized Sackin index, normalized ladder size, and maximum distance from root were calculated for each inferred phylogenetic tree composed of all IgH V genes per each species. (E) The maximum clade credibility (MCC) trees for human IgH V allelic sets. The color of the tip labels indicate identical IgH V gene in both trees.

Figure 4. Laplacian spectra reveals different selective pressures for heavy, kappa, and lambda variable segments. (A) Example spectral densities for the variable heavy chain phylogenies. (B) Heatmap displaying the pairwise Jensen Shannon index based on the Laplacian spectra from each IgH V phylogenetic tree.

Figure 5. Pairwise distances between variable segments of the heavy, light, and kappa immunoglobulin variable gene segments. (A) Pairwise edit distances of the human variable heavy chain for each allelic IgH set. Intensity corresponds to the number of amino acid mutations separating two V gene segments. (B) Distributions of pairwise edit distances of the variable heavy chains for each of the species analyzed. Dashed and solid lines indicate the median and quartiles, respectively. (C) Violin plots of the inter- and intra-species pairwise edit distance for selected species. Dashed and solid lines indicate the median and quartiles, respectively. (D) Mean pairwise distance for each variable heavy chains between species. The intensity represents the median pairwise distance between either two different species (non-diagonal) or the same species (diagonal). (E) The number of mutations separating the same human IgH V genes in the two included allelic sets.

Figure 6. The diversity of the heavy chain repertoire differs by B cell population.

The pairwise edit distance between the top 20 most abundant V genes expressed by four different murine cohorts from an immunization study (Ovalbumin (OVA), 4-hydroxy-3-nitrophenylacetyl conjugated to hen egg lysozyme (NP), no immunization (Control) or hepatitis B virus surface antigen (HBsAg) were calculated and represented as violin plots. The colors represent the different B cell populations (pre B cell IgM, Naïve B cell IgM, bone marrow (BM) IgM plasma cells (PC), BM IgG PCs, splenic IgM PCs, and splenic IgG PCs). The top V genes were calculated using the mean V gene frequencies per cohort ($n \geq 4$ mice).

Figure S1. Number and properties of the germline genes included in this study.

(A) The number of IgH D, IgH J, IgK V, and IgL V genes for each species. Allelic germlines were only included for human IgH V genes. (B) Nucleotide sequence length of the IgH V genes. (C) Amino acid sequence length of the variable light chain genes. (D) Distributions of GC content for IgH V genes annotated either as pseudogenes or as functional alleles from human*01. (E) Distributions of protein instability indices for all IgH V genes and (F) IgK V genes. Dashed and solid lines indicate the median and quartiles, respectively.

Figure S2. Amino acid motifs improve species-separation of variable light chain genes.

Mean amino acid frequencies for variable heavy, kappa, and lambda germline segments. The intensity corresponds to the mean occurrence (%) of a given amino acid across all variable gene segments for each species. (B) t-SNE plot projecting the amino acid frequencies for the individual variable heavy segments into two dimensions. Each triangle represents a single variable kappa chain and diamonds represent variable lambda chain. No allelic variants for human_02 were included in the analysis. (C) t-SNE plot projecting the k-mers ($k=2,3,4$) composition of each variable chain into two dimensions as in B.

Figure S3. Interspecies comparison between variable kappa and lambda chains.

(A-B) The normalized Colless number, normalized Sackin index, normalized ladder size, and maximum distance from root were calculated for each inferred phylogenetic tree composed of all IgK V and IgL V genes per each species, respectively.

Figure S4. Clustering based on Laplacian spectra across variable heavy, kappa, and lambda chains.

Heatmap displaying the pairwise Jensen Shannon index based on the Laplacian spectra from each variable gene phylogenetic tree.

Figure S5. Cross-species phylogenetic tree of variable heavy genes from human, macaque, platypus, and zebrafish.

Sequences from each species are labeled with “H” (blue tip labels), “M” (red tip labels), “P” (black tip labels), and “Z” (green tip labels), respectively.

Figure S6. Pairwise distance from the variable kappa and lambda chains between species normalized by sequence length.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure S1

Figure S2

B

C

Figure S3

Figure S4

Figure S5

Figure S6

Table S1

Human_01	Human_02	Macaque	Mouse	Rat	Platypus	Zebrafish	Rabbit	Alpaca	Dog
M96641_GIHV1-18	X60501_GIHV1-18	BK009022_GIHV1-1	AC090843_GIHV1-11	RatNor_2_chv6_GIHV1-11	NW_001594535_GIHV1-1	BX649502_GIHV1-1	AY386694_GIHV151	AM773729_GIHV1-1	IMGT000001_GIHV1-30
X07448_GIHV1-2	X62104_GIHV1-2	BK009021_GIHV151	AC090843_GIHV1-10	RatNor_2_chv6_GIHV1-10	NW_001594535_GIHV1-10	BX649502_GIHV1-2	AY386694_GIHV1513	AM939697_GIHV152	IMGT000001_GIHV3-10
M96642_GIHV1-24	M96642_GIHV1-24	BK008958_GIHV1510	AC090843_GIHV1-14	RatNor_2_chv6_GIHV1-16	NW_001594535_GIHV1-11	BX649502_GIHV1-3	AY386697_GIHV1517	AM939698_GIHV153	IMGT000001_GIHV3-13
X62109_GIHV1-3	X62107_GIHV1-3	BK008916_GIHV152	AC090843_GIHV1-15	RatNor_2_chv6_GIHV1-18	NW_001594535_GIHV1-12	BX649502_GIHV1-4	AY386697_GIHV1519	AM939699_GIHV154	IMGT000001_GIHV3-16
KF698736_GIHV1-38-4	KF698736_GIHV1-38-4	BK008956_GIHV159	AC090843_GIHV1-16	RatNor_1_chv6_GIHV1-2	NW_001594535_GIHV1-13	AF271897_GIHV1-5	AY386697_GIHV1521	AM939701_GIHV156	IMGT000001_GIHV3-18
X92309_GIHV1-45	X92309_GIHV1-45	BK009149_GIHV2-1	AC090843_GIHV1-17	RatNor_2_chv6_GIHV1-20	NW_001594535_GIHV1-16	BM497502_GIHV10-1	AY386697_GIHV1521-1	AM777379_GIHV3-1	IMGT000001_GIHV3-19
X92343_GIHV1-46	J00240_GIHV1-46	BK009028_GIHV251	AC090843_GIHV1-18	RatNor_3_chv6_GIHV1-22	NW_001594535_GIHV1-17	BX649502_GIHV11-1	AY386697_GIHV1525	AM777379_GIHV3-2	IMGT000001_GIHV3-23
M29809_GIHV1-58	AB019438_GIHV1-58	BK009049_GIHV252	AC073565_GIHV1-19	RatNor_3_chv6_GIHV1-24	NW_001594535_GIHV1-18	BX649502_GIHV11-2	AY386697_GIHV1526	AM777379_GIHV3-3	IMGT000001_GIHV3-25
AB019437_GIHV1-68	M854064B_GIHV1-68	BK009026_GIHV3-10	AC073565_GIHV1-20	RatNor_3_chv6_GIHV1-25	NW_001594535_GIHV1-20	BX649502_GIHV12-2	AY386697_GIHV1528	AM777379_GIHV3-4	IMGT000001_GIHV3-26
L22582_GIHV1-69	Z27506_GIHV1-69	BK009027_GIHV3-11	AC073565_GIHV1-21	RatNor_3_chv6_GIHV1-27	NW_001594535_GIHV1-22	BM495002_GIHV14-1	AY386697_GIHV1529	AM939702_GIHV3514	IMGT000001_GIHV3-28
KF698734_GIHV1-69-2	KF698734_GIHV1-69-2	BK009033_GIHV3-12	AC073565_GIHV1-23	RatNor_3_chv6_GIHV1-29	NW_001594535_GIHV1-23	BX649502_GIHV2-1	AY386697_GIHV1531	AM939707_GIHV3511	IMGT000001_GIHV3-3
KC713934_GIHV1-690	KC713934_GIHV1-690	BK009035_GIHV3-14	AC073565_GIHV1-24	RatNor_3_chv6_GIHV1-30	NW_001594535_GIHV1-24	BX649502_GIHV2-2	AY386697_GIHV1533	AM939708_GIHV3512	IMGT000001_GIHV3-31
M96637_GIHV1-8	HM853452_GIHV1-8	BK009047_GIHV3-20	AC073565_GIHV1-26	RatNor_3_chv6_GIHV1-31	NW_001594535_GIHV1-31	BX649502_GIHV2-3	AY386697_GIHV1534	AM939709_GIHV3513	IMGT000001_GIHV3-32
M96646_GIHV2-26	M6713333_GIHV2-26	BK009048_GIHV3-21	AC073565_GIHV1-27	RatNor_3_chv6_GIHV1-32	NW_001594535_GIHV1-32	BX649502_GIHV3-2	M12180_GIHV1536	AM939710_GIHV3514	IMGT000001_GIHV3-33
X62111_GIHV2-5	KF698731_GIHV2-5	BK009051_GIHV3-22	AC073565_GIHV1-34	RatNor_3_chv6_GIHV1-35	NW_001594535_GIHV1-28	BX649502_GIHV4-1	M602121_GIHV1537	AM939711_GIHV3515	IMGT000001_GIHV3-35
L21969_GIHV2-70	X92241_GIHV2-70	BK009016_GIHV3-5	AC073565_GIHV1-36	RatNor_3_chv6_GIHV1-36	NW_001594535_GIHV1-29	BX649502_GIHV4-2	M93173_GIHV1540	AM939712_GIHV3516	IMGT000001_GIHV3-37
M96652_GIHV3-11	M96652_GIHV3-11	BK009018_GIHV3-6	AC079181_GIHV1-37	RatNor_3_chv6_GIHV1-38	NW_001594535_GIHV1-30	BX649502_GIHV4-3	M93179_GIHV1543	AM939713_GIHV3517	IMGT000001_GIHV3-38
X92217_GIHV3-13	M96653_GIHV3-13	BK009019_GIHV3-7	AC079181_GIHV1-39	RatNor_3_chv6_GIHV1-42	NW_001594535_GIHV1-31	BX649502_GIHV4-5	M93183_GIHV1545	AM939714_GIHV3518	IMGT000001_GIHV3-39
X92216_GIHV3-15	M96654_GIHV3-15	BK009024_GIHV3-9	AC073561_GIHV1-4	RatNor_3_chv6_GIHV1-43	NW_001594535_GIHV1-32	BX649502_GIHV4-6	M93183_GIHV1545	AM939716_GIHV352	IMGT000001_GIHV3-41
M96655_GIHV3-16	AB019440_GIHV3-16	BK009043_GIHV3510	AC079181_GIHV1-42	RatNor_3_chv6_GIHV1-44	NW_001594535_GIHV1-5	BX649502_GIHV4-7	M93185_GIHV1547	AM939718_GIHV3520	IMGT000001_GIHV3-44
M96656_GIHV3-19	M96656_GIHV3-19	BK009044_GIHV3511	AC079181_GIHV1-43	RatNor_3_chv6_GIHV1-47	NW_001594535_GIHV1-6	BX649502_GIHV4-8	AF176997_GIHV1549	AM939721_GIHV3525	IMGT000001_GIHV3-45
M96657_GIHV3-20	KC713937_GIHV3-20	BK008914_GIHV3514	AC079181_GIHV1-47	RatNor_3_chv6_GIHV1-49	NW_001594535_GIHV1-7	BX649502_GIHV4-9	AF176998_GIHV1550	AM939724_GIHV3526	IMGT000001_GIHV3-46
AB019439_GIHV3-21	M96658_GIHV3-21	BK008917_GIHV3515	AC079181_GIHV1-49	RatNor_3_chv6_GIHV1-50	NW_001594535_GIHV1-8	BX649502_GIHV5-1	AF176999_GIHV1551	AM939725_GIHV3527	IMGT000001_GIHV3-47
M96660_GIHV3-23	M35415_GIHV3-23	BK008918_GIHV3516	AC090843_GIHV1-5	RatNor_3_chv6_GIHV1-52	NW_001594535_GIHV1-9	BX649502_GIHV5-2	AF177000_GIHV1552	AM939726_GIHV3528	IMGT000001_GIHV3-5
AC244452_GIHV3-23D	AC244452_GIHV3-23D	BK008923_GIHV3517	AC074329_GIHV1-50	RatNor_3_chv6_GIHV1-53	NW_001627887_GIHV1510	BX649502_GIHV5-3	AF177001_GIHV1553	AM939727_GIHV3529	IMGT000001_GIHV3-50
M83134_GIHV3-30	L26401_GIHV3-30	BK008928_GIHV3518	AC074329_GIHV1-52	RatNor_3_chv6_GIHV1-56	NW_001776005_GIHV1512	BX649502_GIHV5-4	AF177002_GIHV1554	AM939728_GIHV353	IMGT000001_GIHV3-54
AC244456_GIHV3-30-3	M77302_GIHV3-30-3	BK008930_GIHV3519	AC074329_GIHV1-53	RatNor_3_chv6_GIHV1-57	NW_001776005_GIHV1513	BX649502_GIHV5-5	AF177003_GIHV1555	AM939729_GIHV3530	IMGT000001_GIHV3-58
AC244456_GIHV3-30-5	AC245424_GIHV3-30-5	BK008931_GIHV3520	AC074329_GIHV1-54	RatNor_2_chv6_GIHV1-6	NW_001704550_GIHV1514	BX649502_GIHV5-6	AF177004_GIHV1556	AM939730_GIHV3531	IMGT000001_GIHV3-6
AB019439_GIHV3-33	M96665_GIHV3-33	BK008932_GIHV3521	AC074329_GIHV1-55	RatNor_3_chv6_GIHV1-60	NW_001634771_GIHV1515	BX649502_GIHV5-8	AF177005_GIHV1557	AM939731_GIHV3532	IMGT000001_GIHV3-60
M96666_GIHV3-35	M96666_GIHV3-35	BK008933_GIHV3522	AC087166_GIHV1-56	RatNor_3_chv6_GIHV1-62	NW_001634771_GIHV1516	BX649502_GIHV6-1	AF177006_GIHV1558	AM939732_GIHV3533	IMGT000001_GIHV3-61
M96669_GIHV3-38	AB019439_GIHV3-38	BK008934_GIHV3523	AC087166_GIHV1-58	RatNor_3_chv6_GIHV1-64	NW_001627887_GIHV1517	BX649502_GIHV6-2	AF177007_GIHV1559	AM939733_GIHV3534	IMGT000001_GIHV3-67
KF698732_GIHV3-38-3	KF698732_GIHV3-38-3	BK008935_GIHV3525	AC087166_GIHV1-59	RatNor_3_chv6_GIHV1-65	NW_001627887_GIHV1519	BX649502_GIHV7-1	AF177008_GIHV1560	AM939734_GIHV3535	IMGT000001_GIHV3-69
M96672_GIHV3-43	HM853392_GIHV3-43	BK008944_GIHV3526	AC087166_GIHV1-61	RatNor_3_chv6_GIHV1-66	NW_001594535_GIHV2-10	BX649502_GIHV8-1	AF177009_GIHV1561	AM939735_GIHV3536	IMGT000001_GIHV3-7
Z18900_GIHV3-47	AB019438_GIHV3-47	BK008975_GIHV3529	BN0000872_GIHV1-62-1	RatNor_3_chv6_GIHV1-68	NW_001594535_GIHV2-3	BX649502_GIHV8-2	AF177010_GIHV1562	AM939736_GIHV3537	IMGT000001_GIHV3-70
M96675_GIHV3-48	AB019438_GIHV3-48	BK008976_GIHV3530	BN0000872_GIHV1-62-2	RatNor_2_chv6_GIHV1-8	NW_001594535_GIHV2-6	BX649502_GIHV8-3	AF177011_GIHV1563	AM939717_GIHV3538	IMGT000001_GIHV3-75
M96676_GIHV3-49	M96676_GIHV3-49	BK008977_GIHV3531	AC163348_GIHV1-71	RatNor_3_chv6_GIHV10-10	NW_001627887_GIHV2-7	BX649502_GIHV8-4	AF177012_GIHV1564	AM939737_GIHV3539	IMGT000001_GIHV3-76
M96679_GIHV3-53	KF698735_GIHV3-53	BK008987_GIHV3535	AC079339_GIHV1-63	RatNor_3_chv6_GIHV10-11	NW_001594535_GIHV3-1	BX649502_GIHV9-1	AF177013_GIHV1565	AM939738_GIHV354	IMGT000001_GIHV3-8
M96682_GIHV3-64	AB019437_GIHV3-64	BK009015_GIHV354	AC079339_GIHV1-64	RatNor_3_chv6_GIHV10-15	NW_001594535_GIHV3-3	BX649502_GIHV9-2	AF177014_GIHV1566	AM939739_GIHV3540	IMGT000001_GIHV3-80
X92218_GIHV3-66	Z27506_GIHV3-66	BK008995_GIHV3540	AC079339_GIHV1-66	RatNor_3_chv6_GIHV10-18	NW_001594535_GIHV3-4	BX649502_GIHV9-3	AF177015_GIHV1567	AM939740_GIHV3541	IMGT000001_GIHV3-81
AJ87484A_GIHV3-69-1	AJ87484B_GIHV3-69-1	BK008996_GIHV3541	AC079339_GIHV2-67	RatNor_3_chv6_GIHV10-19	NW_001594535_GIHV3-5	BX649502_GIHV9-4	AF177016_GIHV1568	AM939741_GIHV3542	IMGT000001_GIHV3-82
M96649_GIHV3-7	X92288_GIHV3-7	BK008997_GIHV3542	AC079339_GIHV1-69	RatNor_2_chv6_GIHV10-2	NW_001594535_GIHV5-1				
X92206_GIHV3-72	Z29979_GIHV3-72	BK008998_GIHV3543	AC090843_GIHV1-7	RatNor_2_chv6_GIHV10-4					
X70197_GIHV3-73	AB019437_GIHV3-73	BK009000_GIHV3544	AC163348_GIHV1-71	RatNor_2_chv6_GIHV10-5					
L33851_GIHV3-74	Z17392_GIHV3-74	BK009002_GIHV3545	AC163348_GIHV1-72	RatNor_2_chv6_GIHV10-7					
M96651_GIHV3-9	HM855577_GIHV3-9	BK009004_GIHV3546	AC160473_GIHV1-74	RatNor_2_chv6_GIHV10-8					
HM855939_GIHV3-1L1	HM855939_GIHV3-1L1	BK009060_GIHV3548	AC160473_GIHV1-75	RatNor_1_chv6_GIHV11-1					
X02714_GIHV4-28	M83113_GIHV4-28	BK009020_GIHV355	AC160473_GIHV1-76	RatNor_2_chv6_GIHV11-10					
L10080_GIHV4-30-2	M95122_GIHV4-30-2	BK009041_GIHV359	AC160473_GIHV1-77	RatNor_2_chv6_GIHV11-12					
Z14238_GIHV4-30-4	Z14239_GIHV4-30-4	BK009029_GIHV4-2	AC160473_GIHV1-78	RatNor_2_chv6_GIHV11-2					
L10098_GIHV4-31	M96643_GIHV4-31	BK008950_GIHV4510	AC160990_GIHV1-80	RatNor_2_chv6_GIHV11-3					
AB019439_GIHV4-34	M96644_GIHV4-34	BK008961_GIHV4511	L60990_GIHV1-81	RatNor_2_chv6_GIHV11-4					
Z12367_GIHV4-38-2	AC233755_GIHV4-38-2	BK008966_GIHV4512	AC160990_GIHV1-82	RatNor_2_chv6_GIHV11-8					
AB019439_GIHV4-39	X05715_GIHV4-39	BK008977_GIHV4513	AC160990_GIHV1-84	RatNor_2_chv6_GIHV11-9					
X02713_GIHV4-4	X02322_GIHV4-4	BK009026_GIHV4514	AC160990_GIHV1-85	RatNor_2_chv6_GIHV12-2					
AB019438_GIHV4-59	M29812_GIHV4-59	BK008945_GIHV4515	AC090843_GIHV1-9	RatNor_6_chv6_GIHV2-1					
M29811_GIHV4-61	L10097_GIHV4-61	BK008955_GIHV4516	AC073561_GIHV10-1	RatNor_1_chv6_GIHV2-13					
X92227_GIHV5-10-1	X92279_GIHV5-10-1	BK008964_GIHV4517	AC073561_GIHV10-3	RatNor_1_chv6_GIHV2-15					
M96686_GIHV5-51	M18806_GIHV5-51	BK008962_GIHV4518	AC073563_GIHV11-1	RatNor_1_chv6_GIHV2-16					
X92213_GIHV5-78	X92213_GIHV5-78	BK008971_GIHV4519	AC073563_GIHV11-2	RatNor_1_chv6_GIHV2-19					
J04097_GIHV6-1	Z14223_GIHV6-1	BK009034_GIHV452	AC073590_GIHV12-3	RatNor_1_chv6_GIHV2-27					
IMGT000035_GIHV7-4-1	X62110_GIHV7-4-1	BK008911_GIHV455	AC073589_GIHV13-1	RatNor_1_chv6_GIHV2-30					
AB019437_GIHV7-81	AB019437_GIHV7-81	BK008920_GIHV456	AC073590_GIHV13-2	RatNor_1_chv6_GIHV2-32					
IMGT000035_GIHV8-51-1	IMGT000035_GIHV8-51-1	BK008938_GIHV457	AC079273_GIHV14-1	RatNor_1_chv6_GIHV2-38					
			BN0000872_GIHV2-6-8	RatNor_1_chv6_GIHV2-39					
		BK008949_GIHV459	AC073563_GIHV14-2	RatNor_1_chv6_GIHV2-39					
		BK009012_GIHV5-1	AC073563_GIHV14-3	RatNor_6_chv6_GIHV2-4					
		BK009040_GIHV5-2	AC073589_GIHV14-4	RatNor_1_chv6_GIHV2-40					
		BK008990_GIHV551	AC090843_GIHV15-2	RatNor_1_chv6_GIHV2-41					
		BK009057_GIHV6-1	AC073563_GIHV16-1	RatNor_1_chv6_GIHV2-43					
		X02714_GIHV7-1	L33944_GIHV15100	L33944_GIHV15100					
		BK008935_GIHV751	L33942_GIHV15103	L33942_GIHV15103					
			L33946_GIHV15107	L33946_GIHV15107					
			L33947_GIHV15108	L33947_GIHV15108					
			L00535_GIHV155	L00535_GIHV155					
			L26856_GIHV1565	RatNor_1_chv6_GIHV2-55					
			L26858_GIHV1567	RatNor_1_chv6_GIHV2-59					
			L26860_GIHV1568	RatNor_6_chv6_GIHV2-6					
			L26925_GIHV1587	RatNor_1_chv6_GIHV2-61					
			L26931_GIHV1592	RatNor_1_chv6_GIHV2-63					
			L26934_GIHV1595	RatNor_1_chv6_GIHV2-64					
			L26935_GIHV1596	RatNor_1_chv6_GIHV2-65					
			AC090887_GIHV2-2	RatNor_1_chv6_GIHV2-70					
			AC090887_GIHV2-3	RatNor_1_chv6_GIHV2-72					
			AC090887_GIHV2-4	RatNor_1_chv6_GIHV2-75					
			AC090887_GIHV2-5	RatNor_1_chv6_GIHV2-76					
			CAA01078923_GIHV2-6	RatNor_1_chv6_GIHV2-77					

Table S2

Species	Kappa	Lambda
Human	Homo_sapiens-ABBA01-ABBA01165617.1	Homo_sapiens-ABBA01-ABBA01055078.1
Macaque	Macaca_fascicularis-AQIA01-AQIA01017094.1	Macaca_fascicularis-AQIA01-AQIA01007566.1
Mouse	Mus_musculus-AAHY01-AAHY01054000.1	Mus_musculus-AAHY01-AAHY01126128.1
Rat	Rattus_norvegicus-AABR07-AABR07051741.1	Rattus_norvegicus-AABR07-AABR07034730.1
Platypus	Ornithorhynchus_anatinus-AAPN01-AAPN01029322.1	Ornithorhynchus_anatinus-AAPN01-AAPN01173604.1
Zebrafish	Danio_rerio-CABZ01-CABZ01052991.1	Danio_rerio-CABZ01-CABZ01038528.1
Rabbit	Oryctolagus_cuniculus-AAGW02-AAGW02007316.1	Oryctolagus_cuniculus-AAGW02-AAGW02067712.1
Alpaca	Vicugna_pacos-JEMW01-JEMW01023803.1	Vicugna_pacos-JEMW01-JEMW01002007.1
Dog	Canis_lupus_familiaris-AAEX03-AAEX03010921.1	Canis_lupus_familiaris-AAEX03-AAEX03014800.1