Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans

Shishi Luo[1,2], Jane A. Yu[1], Heng Li[3], Yun S. Song[1,2,4,5,*]

[1] Computer Science Division, University of California, Berkeley, Berkeley, CA 94720, USA

[2] Department of Statistics, University of California, Berkeley, Berkeley, CA 94720, USA

[3] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[4] Departments of Mathematics and Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

[5] Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

[*] To whom correspondence should be addressed: yss@berkeley.edu

Abstract

Despite the crucial role played by the immunoglobulin heavy variable (IGHV) and T cell beta variable (TRBV) loci in adaptive immune function, their genetic variation in the human population remains poorly characterized. Generated through a process of gene duplication/deletion and diversification, these loci can vary extensively between individuals in copy number and contain genes that are highly similar. These characteristics make the identification and analysis of these loci technically challenging. Here, we present a comprehensive study of the functional gene segments in the IGHV and TRBV loci, quantifying their copy number and single nucleotide variation in a globally diverse sample of 109 (IGHV) and 286 (TRBV) humans. We find that despite the shared molecular and functional characteristics between the IGHV and TRBV gene families, they exhibit starkly different patterns of variation. In particular, we estimate that there are hundreds of copy number haplotypes in the IGHV locus (instances that have differences in the number of functional gene segments), while the TRBV locus has only a few copy number haplotypes. We also find that the TRBV locus has a greater or at least equal propensity to mutate, as evidenced by greater single nucleotide variation, compared to the IGHV locus. This is consistent with the observation that, across multiple species, the TRBV gene family is more diverse than the IGHV gene family. As well as suggesting that the IGHV and TRBV loci have taken different paths over their evolutionary history, these results also indicate that adaptive immune repertoire sequencing and analysis may be much more challenging for immunoglobulins than T cell receptors.

Introduction

By some estimates, genomic variation due to copy number differences underlies more variation in the human genome than that due to single nucleotide differences[1]. Yet copy number variation remains challenging to quantify and analyze. Nowhere is this more true than in genomic regions that contain gene families: collections of genes formed through the process of duplication/deletion and diversification of contiguous stretches of DNA[2]. Two gene families that are of particular biomedical relevance but for which variation is not well characterized are the immunoglobulin heavy variable family (IGHV), a 1 Mb locus located on chromosome 14[3,4], and the T cell receptor beta variable family (TRBV), a 500 kb locus located on chromosome 7[5]. Both loci undergo VDJ recombination, providing the V (variable) component in the biosynthesis of adaptive immune receptors: the IGHV for the heavy chain of the B cell receptor and the TRBV for the beta chain of the T cell receptor[6]. In the human genome, both loci are organized as a series of approximately 45 functional V gene segments and are adjacent to a collection of D (diversity) and J (joining) segments. Both loci are present in the genomes of all vertebrates known to have an adaptive immune system, although the arrangement of the IGHV locus can differ between species[7–9]. Indeed, the genes comprising the IGHV and TRBV loci are distant homologs, and are believed to derive from a common ancestral locus in a vertebrate contemporaneous with or predating jawed fishes. That these two loci share genomic features and evolutionary origins makes them an ideal system for a comparative study in gene family evolution.

Here, we present a quantitative catalog of copy number and allelic variation in the IGHV and TRBV loci, by applying a customized genotyping pipeline (based on reference [10]) to data from the Simons Genome Diversity Project[11], which performed whole-genome sequencing of a globally diverse sample of human individuals from over a hundred populations. Such characterization of genetic variation in the immune receptor loci is much needed in the burgeoning field of computational immunology[12,13], where the relative abundances of germline variants will help in determining clonal lineages from VDJ sequences.

Results

A brief note about gene nomenclature: for the bioinformatic analysis, it was necessary to group together gene segments that were operationally indistinguishable but which had distinct names because they occupy physically different positions on the genome. Our departure from this standard nomenclature is detailed in the Methods and Supplementary Note and is also explained where needed below.

To minimize confusion around terminology, we use *polymorphism* as a general term for a genomic unit (nucleotide position or gene segment) that exhibits variation between genomes. Different instances of a particular polymorphism are called *variants*, e.g. a single nucleotide variant or a gene copy number variant. The term *allele* is reserved exclusively for referring to variants of a gene, as in the allele IGHV1-69*01, which is a gene-length variant of the IGHV1-69 gene segment and which may differ in more than a single nucleotide from other alleles of IGHV1-69. This is in line with the usage in the immunogenetics community. We use *haplotype* to refer to the set of operationally distinguishable gene segments that are inherited from a single parent. For example, an individual has two haplotypes of the IGHV locus with one haplotype carrying, say, one copy of IGHV1-69 and the other carrying two copies, say IGHV1-69 and IGHV1-69D.

General patterns of variation

Summary statistics from our quantification of copy number and single nucleotide variation paint a fascinating picture of how the IGHV and TRBV gene families — in spite of their shared evolutionary origin, similar genomic arrangement, and a common VDJ rearrangement mechanism— have traveled down starkly different evolutionary paths (Figure 1). The evolutionary process of gene duplication/deletion appears to have been faster in the IGHV locus than in the TRBV locus. This is evident in the greater variation in the number of operationally distinguishable IGHV gene segments than in TRBV gene segments (Figure 1A). In contrast, the process of diversification — as measured by the number of single nucleotide polymorphisms per gene segment — seems to be comparable or possibly faster in the TRBV locus than in the IGHV locus (Figure 1B). To be conservative, this single nucleotide

polymorphism analysis was restricted to segments which are predominantly two copies per individual and which do not share subsequences with other segments.

With these different evolutionary dynamics, we expect the IGHV gene family to comprise genes that are more similar to each other on average than the TRBV gene family. Indeed, this is precisely the pattern we see in the human IGHV and TRBV gene families and across twelve other vertebrate species, including four primates, six non-primate mammals, one reptile, and one fish (Figure 1C, Methods). We also observe less homology between species for the IGHV gene family compared to the TRBV family (Supplementary Figure 2). This is consistent with our finding that gene duplication and deletion occurs more frequently in the IGHV locus: IGHV homologs that are shared between species are erased more frequently than TRBV homologs. It may therefore also be the case that in these other vertebrate species, the TRBV locus undergoes less frequent rapid gene duplication/deletions and more frequent nucleotide mutations than the IGHV locus.

Copy number variation

Figure 2 and Figure 3 provide the relative abundances of variants for some common copy number polymorphisms in the IGHV and TRBV loci[4,14–18]. This, and the more detailed Supplementary Figures 3 and 4, should serve as a useful reference for the computational immunology community.

Lack of geographical associations. In the majority of cases, we find that the distribution of variants within a geographic region is consistent with the global distribution (Supplementary Figures 5 and 6). The two exceptions are: (i) the polymorphism involving IGHV1-69, where the duplication/insertion variant is the major variant among individuals from Africa, despite being a minor variant (28%) of the global sample, and (ii) the three-gene deletion of TRBV5-8, TRBV7-8, and TRBV6-9, which is the major variant among Native Americans, but appears in only 5% of our sample globally. In neither of these two cases is there evidence to suggest the absence of any particular gene is fatal. Overall, these indicate that IGHV and TRBV copy number variation was established early in the history of *homo sapiens*, and that it is unlikely

that the presence of particular IGHV or TRBV gene segments is vital against any region-specific pathogens.

No correlation between copy number polymorphisms. We find effectively no correlation between copy number polymorphisms in either IGHV or TRBV (Supplementary Figures 7 and 8). The average value of R^2, the square of the Pearson correlation coefficient, between segments in the different polymorphisms shown in Figure 2 and Figure 3 is 0.021 for the IGHV gene segments and 0.004 for the TRBV gene segments. Thus, the polymorphisms are essentially independent, and we can estimate the number of copy number haplotypes in the two loci. With three polymorphisms each with 2 (haploid) variants, and with the set {IGHV3-30, IGHV3-30-3, IGHV3-30-5, IGHV3-33} and {IGHV3-23, IGHV3-23D} exhibiting an estimated 7 and 4 (haploid) copy number variants, respectively, this gives approximately 200 IGHV haplotypes (2 x 2 x 2 x 7 x 4), assuming independence between the common copy number polymorphisms. The analogous calculation for TRBV leads to about a handful of haplotypes (2 x 2).

This has practical implications for germline IGHV haplotyping: approaches for cataloging variation by sequencing the one megabase locus in full[4,19] will need to consider the possibility that even in a sample of hundreds of individuals, there will be copy number differences between a substantial fraction of haplotypes. Indeed, we find that when we draw two individuals at random from our sample, there is a 98% chance that they will have different sets of IGHV segments present or absent, but only an 11% chance they will have different sets of TRBV segments present or absent. This calculation is based on a coarser, more robust measure of copy number haplotype, where we identify each individual by the presence or absence of a segment, and is therefore conservative. These numbers remain approximately the same even when we restrict our comparisons to individuals within geographical regions, again indicating that the presence or absence of functional segments does not segregate by geographic region (Supplementary Table 1). These results provide quantitative support for the conjecture made in [16] that "no chromosomes contain the same set of V_H gene segments," where V_H refers to IGHV.

Our results are also of immediate relevance to the adaptive immune receptor repertoire sequencing community. The greater complexity in the IGHV locus suggests that using data analysis methods interchangeably between T cell receptor sequences and B cell receptor sequences may not be optimal. The majority of TRBV genes are operationally distinguishable and appear as a single copy per haplotype. Since T cell receptors do not undergo further somatic hypermutation, it makes sense to construct so-called "public" T cell receptor repertoires and analyze individual repertoires in relation to common public repertoires. In contrast, the majority of IGHV genes either vary in copy number or share long subsequences in common with other genes/pseudogenes/orphon genes in the IGHV family (Supplementary Figure 3). Furthermore, immunoglobulins undergo genetic modification via somatic mutation. The analysis of the antibody repertoire may therefore need to be customized to each individual.

The mechanisms underlying the higher rate of gene duplication/deletion in IGHV remain to be ascertained. In the process of generating copy number estimates, we found that 60% of IGHV segments share long subsequences (>75 bp) with other IGHV genes/pseudogenes/orphon genes while only 23% of TRBV segments did (Supplementary Figures 3 and 4, Supplementary Note). This would be consistent with ectopic gene conversion occurring frequently in the IGHV locus, especially since ectopic gene conversion is crucial to the biosynthesis of immunoglobulins in animals such as chickens and rabbits, which have gut-associated lymphoid tissue[20]. Closer proximity to the telomere and higher abundance of genome-wide interspersed repeats, also known as transposable elements, are also associated with structural variation[21]. In this respect, the B cell receptor locus, located at the end of the telomere on chromosome 14 and with genome-wide interspersed repeats making up 42% of its nucleotide content[3], would be expected to exhibit greater structural variation than the T cell receptor locus, which is located 15 million base pairs interior to the telomere of chromosome 7 and has 30% of its nucleotide content as genome-wide interspersed repeats[5].

Single nucleotide and allelic variation

Because copy number variation in the IGHV and TRBV loci has not been well characterized until now, it has not been straightforward to compare the single nucleotide and allelic

variation in gene segments between the two loci. A gene segment with higher copy number could be perceived as exhibiting greater single nucleotide or allelic variation, even though it experiences the same rate of per-base substitution. This is because a duplicate with mutations is difficult to distinguish from a true allele. Having quantified copy number variation of genes across the two loci, we are able to perform this comparison while minimizing the confounding factor of copy number variation. Specifically, we compare single nucleotide and allelic variation in IGHV and TRBV gene segments that are in two copies in the vast majority of individuals in our sample and for which there is minimal read-mapping ambiguity (11 such IGHV segments, 41 TRBV segments, see Supplementary Figures 3 and 4). We will refer to such gene segments as "two-copy" for short. Several segments in TRBV and in IGHV occur as multi-copies, and some segments have single nucleotide polymorphisms with more than two variants. Our pipeline does not handle calling alleles or single nucleotide variants for these more complex cases.

IGHV and TRBV have comparable rates of nucleotide substitution. We find that when restricted to the set of two-copy gene segments in IGHV and TRBV, the two loci have comparable summary measures of single nucleotide and allelic variation (Table 1). If anything, the TRBV gene segments exhibit greater single nucleotide and allelic diversity. This is seemingly surprising, because if we calculate the average number of alleles per gene segment (as per the IMGT database[10]), without regard to the segment's copy number, there are 5 alleles per operationally distinguishable IGHV gene segment compared to 2 alleles per operationally distinguishable TRBV gene segment. The discrepancy in the two measures of allelic variation does not seem to be due to any under-representation of TRBV alleles in the IMGT database: the fraction of putative novel alleles called in our sample is similar between the IGHV and TRBV gene segments (Table 1, third row). As a reference for the antibody repertoire sequencing community, we have provided the relative abundances of alleles for the two-copy gene segments calculated from our sample in Supplementary Tables 2 and 3.

Putatively novel alleles. In addition to restricting our single nucleotide and allelic analysis to two-copy gene segments, we called allele or single nucleotide variant only if a variant is present in two or more individuals. With this criteria, we called 29 IGHV alleles, of which 13

are putatively novel and 105 TRBV alleles, of which 47 are putatively novel. Of these novel alleles, it is notable that 5 IGHV alleles and 13 TRBV alleles appeared at least 10 times in our sample (we count homozygous alleles as appearing twice; Supplementary Table 4). Some of these alleles like IGHV1-45*02_g123A, TRBV10-1*02_g234T, and TRBV12-5*01_c27G are present in high frequency across all geographic regions. That these novel variants are comprehensively present is further evidence that existing databases of IGHV and TRBV alleles are not yet complete.

Single nucleotide/Allelic variants private to geographic regions. We found 5 single nucleotide variants in IGHV gene segments that are private to a single geographic region, and 26 such variants in TRBV gene segments (Supplementary Table 5). These variants are not rare: a majority of them are present at greater than 10% frequency in geographic region they are private to, in one case the frequency is 42%. For both loci, the geographic region of Africa had a disproportionate share of such variants: of the five IGHV single nucleotide variants that were private to a geographic region, all five were private to Africa, and of the 26 single nucleotide variants exclusive to a region for TRBV, 20 (77%) were private to Africa. This particular feature of samples from the Africa region is also apparent in our allelic variation analysis (Supplementary Table 6). Of the 29 IGHV alleles we called, 5 out of 5 private alleles were private to Africa. Similarly, of the 98 TRBV alleles we called, 15 out of 21 (71.4%) private alleles were private to Africa. These findings of higher levels of diversity primarily in Africa are consistent with prior studies [22] and with the percentage of exon-located single nucleotide polymorphisms that are private to Africa across the entire genome (72.1%). Such variants, which are private to a particular region, may be promising candidates for investigating genetic variants that are beneficial against infectious diseases endemic in a geographic region. For a complete table of alleles and single nucleotide variants private to a particular region, see Supplementary Table 6 and Supplementary Table 5.

Methods

Full methodological details are provided in Supplementary Note.

Gene nomenclature. The following sets of gene segments were considered operationally indistinguishable for our bioinformatic analysis: {IGHV3-23, IGHV3-23D}, {IGHV3-30, IGHV3-30-3, IGHV3-30-5, IGHV3-33}, {IGHV3-53, IGHV3-66}, {IGHV3-64, IGHV3-64D}, {IGHV1-69, IGHV1-69D}, {IGHV2-70, IGHV2-70D}, {TRBV4-2, TRBV4-3}, {TRBV6-2, TRBV6-3}, {TRBV12-3 TRBV12-4}.

SGDP Dataset. Paired reads, 100 bp in length, from 286 individuals in the Simons Genome Diversity Project[11] were mapped to the set of functional alleles (IGHV or TRBV) as per the IMGT database[10]. Of the 286 individuals, only those from non-cell line, i.e. blood and saliva DNA sources (109 in total), could be used for IGHV analysis. This is because in these cell lines, which are based on immortalized B cells, the IGHV locus is truncated relative to germline configuration due to VDJ recombination. Details of individual samples can be found in Supplementary Data Table 1 of reference [11].

Copy number and allele calls. For each of the IGHV and TRBV loci, we filter the set of raw reads, aiming to minimize reads that have been erroneously mapped to a functional gene segment. This required taking into account idiosyncrasies of individual segments, especially their similarity to pseudogenes and orphon genes. De novo assembly is then applied to the reads for each segment in each individual. The contig coverage calculated from the *de novo* assembler, after appropriate scaling, becomes our estimate of copy number. The contig and reads are then used to call single nucleotide polymorphisms and to phase alleles. We only called alleles for two-copy segments: those that occurred as single copy in the locus in the vast majority of individuals in our dataset and which do not share substantial nucleotide sequence with any other gene segments.

Other species. Amino acid sequences for IGHV segments and TRBV segments for the interspecies comparison were obtained from vgenereportoire.org[23]. Species are: *homo sapiens* (human), *pan troglodytes* (chimpanzee), *gorilla gorilla* (gorilla), *pongo abelii* (orangutan),

macaca mulatta (rhesus macaque), *mus musculus* (mouse), *canis lupus familiaris* (dog), *oryctolagus cuniculus* (rabbit), *orcinus orca* (orca), *monodelphis domestica* (opossum), *ornithorhynchus anatinus* (platypus), *crocodylus porosus* (crocodile), *danio rerio* (zebrafish).

Acknowledgments

We gratefully acknowledge David Reich for useful discussions and for making short-read data from the Simons Genome Diversity Project available to us. This research is supported in part by an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering. YSS is a Chan Zuckerberg Biohub investigator.

References

1. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
2. Nei, M. & Rooney, A. P. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu. Rev. Genet.* **39,** 121–152 (2005).
3. Matsuda, F. *et al.* The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. *J. Exp. Med.* **188,** 2151–2162 (1998).
4. Watson, C. T. *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.* **92,** 530–546 (2013).
5. Rowen, L., Koop, B. F. & Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Sci. Wash.* **272,** 1755 (1996).
6. Murphy, K. & Weaver, C. *Janeway's Immunobiology, 9th edition*. (Garland Science, 2016).
7. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11,** 47–59 (2010).
8. Das, S., Nozawa, M., Klein, J. & Nei, M. Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates. *Immunogenetics* **60,** 47–55 (2008).
9. Cannon, J. P., Haire, R. N., Rast, J. P. & Litman, G. W. The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. *Immunol. Rev.* **200,** 12–22 (2004).
10. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* **33,** D256–D261 (2005).

11. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538,** 201–206 (2016).
12. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32,** 158–168 (2014).
13. Robins, H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.* **25,** 646–652 (2013).
14. Pramanik, S. *et al.* Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. *BMC Genomics* **12,** 78 (2011).
15. Chimge, N.-O. *et al.* Determination of gene organization in the human IGHV region on single chromosomes. *Genes Immun.* **6,** 186–193 (2005).
16. Li, H., Cui, X., Pramanik, S. & Chimge, N.-O. Genetic diversity of the human immunoglobulin heavy chain VH region. *Immunol. Rev.* **190,** 53–68 (2002).
17. Seboun, E., Houghton, L., Hatem, C. J., Lincoln, R. & Hauser, S. L. Unusual organization of the human T-cell receptor beta-chain gene complex is linked to recombination hotspots. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 5026–5029 (1993).
18. Brennan, R. M. *et al.* The Impact of a Large and Frequent Deletion in the Human TCR β Locus on Antiviral Immunity. *J. Immunol.* **188,** 2742–2748 (2012).
19. Steele, E. J. & Lloyd, S. S. Soma-to-germline feedback is implied by the extreme polymorphism at IGHV relative to MHC. *BioEssays* **37,** 557–569 (2015).
20. Flajnik, M. F. Comparative analyses of immunoglobulin genes: surprises and portents. *Nat. Rev. Immunol.* **2,** 688–698 (2002).
21. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8,** 272–285 (2007).

22. Brzezinski, J. L., Deka, R., Menon, A. G., Glass, D. N. & Choi, E. Variability in TRBV haplotype frequency and composition in Caucasian, African American, Western African and Chinese populations. *Int. J. Immunogenet.* **32,** 413–420 (2005).

23. Olivieri, D., Faro, J., von Haeften, B., Sánchez-Espinel, C. & Gambón-Deza, F. An automated algorithm for extracting functional immunologic V-genes from genomes in jawed vertebrates. *Immunogenetics* **65,** 691–702 (2013).

24. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. *PLOS ONE* **8,** e82138 (2013).

25. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).

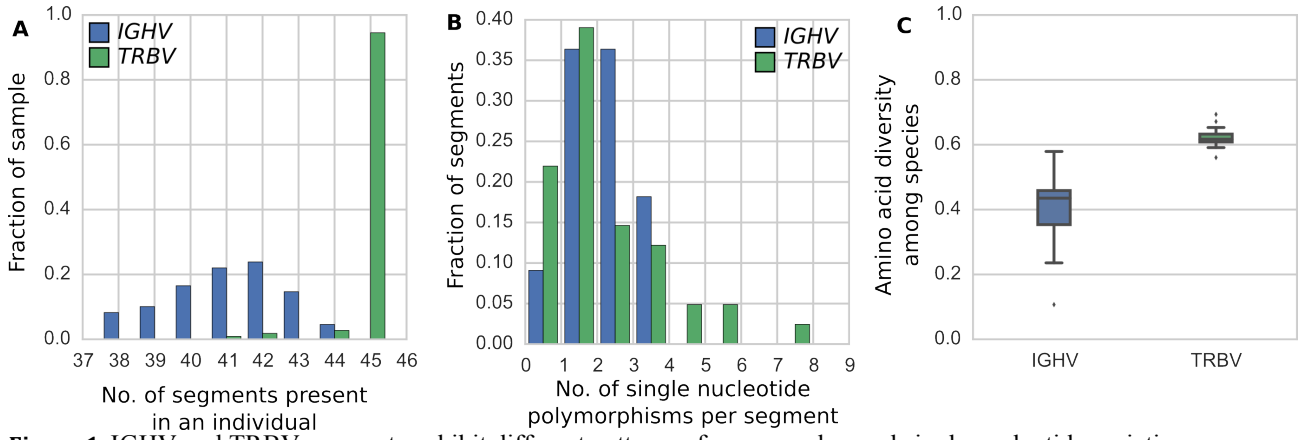

Figure 1. IGHV and TRBV segments exhibit different patterns of copy number and single nucleotide variation, consistent with the genetic diversity in these gene families observed across species. (A) The number of distinct IGHV (blue) and TRBV (green) segments present in each of the 109 individuals from blood and saliva samples. Supplementary Figure 1 shows (A) for TRBV in the full set of 286 individuals. (B) The number of single nucleotide polymorphisms per two-copy IGHV (blue) and TRBV (green) segment. We report a single nucleotide polymorphism only if it is found in at least two out of the 109 genomes from blood and saliva samples. The two distributions are not statistically significantly different (p value of two-sample Kolmogorov-Smirnov test between the blue and green distribution is 0.97). (C) Box plots for average pairwise diversity between IGHV segments and TRBV segments in each of thirteen vertebrate species. For each species, pairwise alignments of all pairs of IGHV segments and all pairs of TRBV segments were performed using ssw [24], an implementation of the Smith-Waterman algorithm [25].

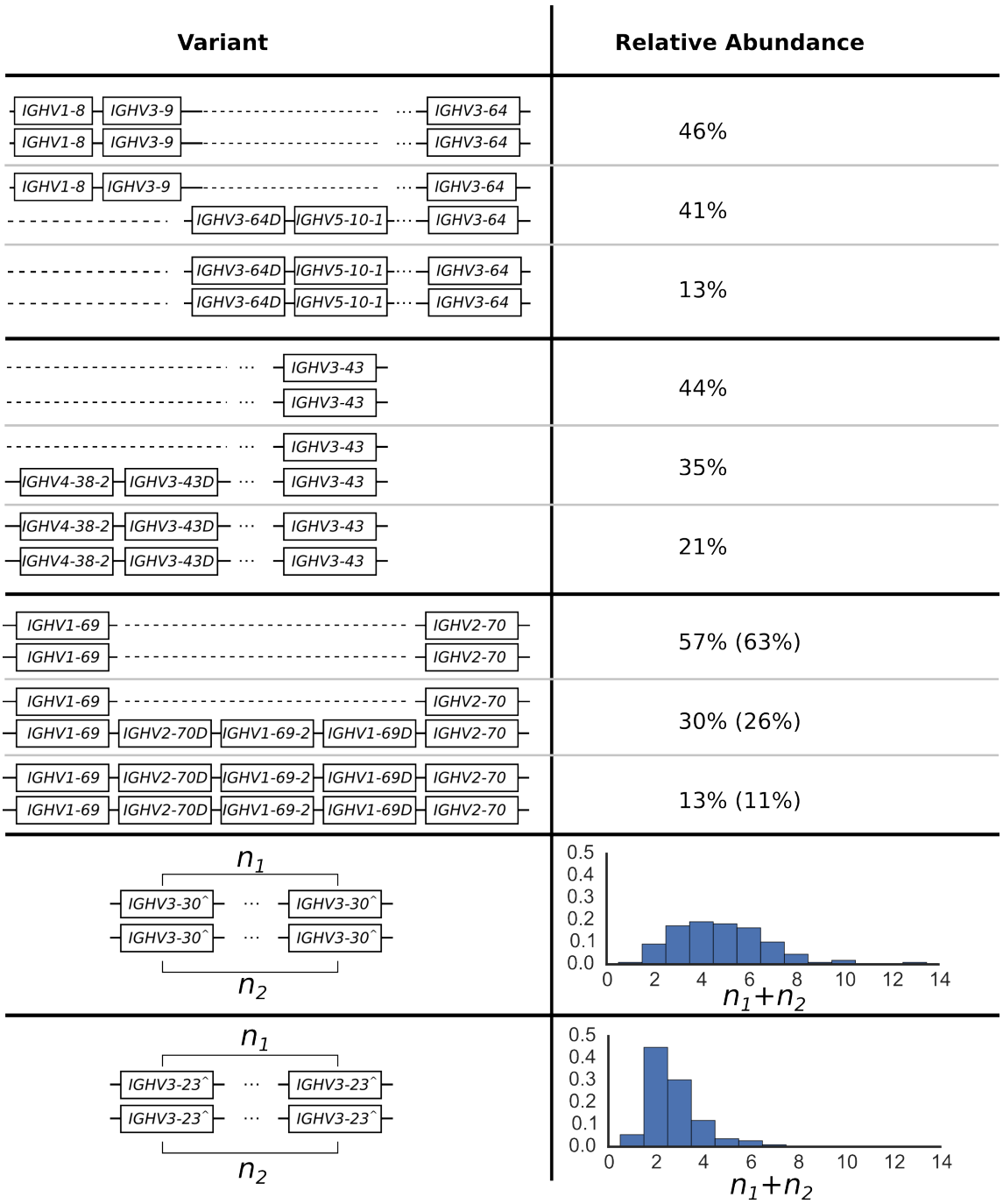

Figure 2. The distribution of common IGHV copy number polymorphisms. Schematics in the left column show the genomic configuration of these polymorphisms determined by previous studies [4]. The right column displays the relative abundance in the sample of 109 individuals. For the polymorphism involving IGHV1-69, we also show the relative abundances in the full sample of 286 individuals in parentheses. This is because the impact of VDJ recombination on cell line samples in this J-distal part of the locus has a negligible effect on our copy number estimates. Note that we use IGHV3-30^ as shorthand for the set {IGHV3-30, IGHV3-30-3, IGHV3-30-5, IGHV3-33} and IGHV3-23^ for {IGHV3-23, IGHV3-23D}. The relative abundances for the copy number variants of IGHV3-30^ and IGHV3-23^ are for the total number summed over the two haplotypes in an individual. We estimate the largest haploid number to be 7 for IGHV3-30^ (because the smallest two numbers to sum to 13, the largest copy number called, are 6 and 7) and 4 for IGHV3-23^ (likewise, 4 and 3 are the smallest two numbers to sum to 7).

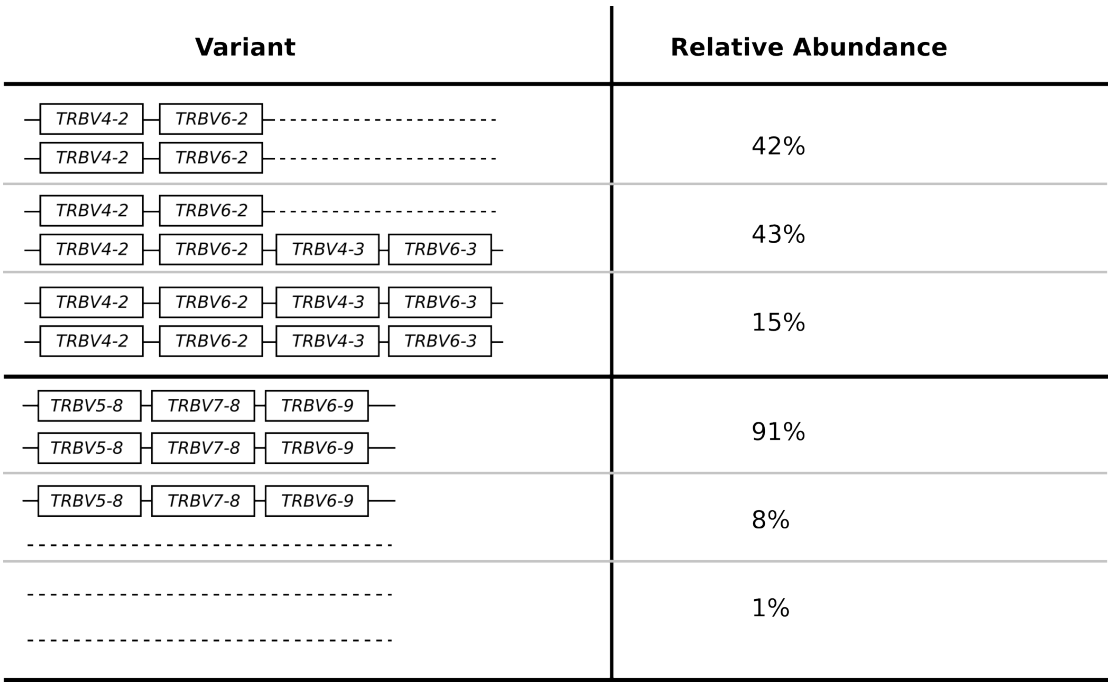

Figure 3. The distribution of common TRBV copy number polymorphisms. Schematics in the left column show putative genomic configurations of the polymorphisms. The right column displays the relative abundance in the full sample of 286 individuals.

Table 1. Summary statistics for single nucleotide and allelic variation in IGHV and TRBV. The results tabulated are computed using the same set of 109 individuals and are restricted to the two-copy segments described in the text.

	IGHV	TRBV
Average bp difference per pairs of alleles	4.6%	5.1%
Average # of SNPs per segment	1.8	2.1
Fraction of novel alleles out of all observed alleles	45%	45%