Computational Tools for Immune Repertoire Characterization and Primer Set Design

Jane Yu

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-1
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-1.html

January 2, 2020

Computational Tools for Immune Repertoire Characterization and Primer Set Design

by

Jane Yu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Haiyan Huang
Assistant Professor Nir Yosef

Fall 2019

Computational Tools for Immune Repertoire Characterization and Primer Set Design

Abstract

Computational Tools for Immune Repertoire Characterization and Primer Set Design

by

Jane Yu

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Yun S. Song, Chair

The enormous decrease in the cost of genomic sequencing over the past two decades has enabled researchers to revisit previously unaddressable questions in sequence analysis. However, this boom of genomic information has introduced new sets of problems that often demand computationally efficient methods. In this work, we describe computational tools for two such settings involving large-scale genomic data: 1) estimating copy number and allelic variation in two highly complex gene families, and 2) selective sequencing of a target genome in a complex DNA sample.

We first describe a method that takes short reads from high-throughput sequencing and characterizes both copy number and allelic variation in the IGHV and TRBV loci. These two loci can vary extensively between individuals in copy number and contain genes that are highly similar, making their analysis technically challenging. Additionally, we have conducted the first study of a globally diverse sample of hundreds of individuals in these two loci from over a hundred populations. In addition to providing insight into the different evolutionary paths of the IGHV and TRBV loci, our results are also important to the adaptive immune repertoire sequencing community, where the lack of frequencies of common alleles and copy number variants is hampering existing analytical pipelines.

In our second problem setting, we describe SOAPswga, an optimized and parallelized pipeline for primer design in the context of selective amplification. Unlike previous heuristic-based methods, SOAPswga uses machine learning methods to evaluate both individual primers and primer sets. Additionally, rather than brute force search for primer sets, such as in predecessor methods, SOAPswga uses branch-and-bound principles to pursue only the most promising sets. These optimizations, including the parallelization of each step, allow for a huge decrease in runtime from the order of weeks to minutes. We also discuss the results of our pipeline applied to the selective amplification of *Mycobacterium tuberculosis* in a sample of human blood. Lastly, we expand on the importance of this work, and in general, its potential usefulness to any setting consisting of targeted sequencing.

To Mommia and Daddy

Contents

List of Figures

List of Tables

Acknowledgments

My advisor, Yun, is one of the most devoted and hard-working individuals I have ever met. I could always depend on Yun to make time in his extremely busy schedule for me and I'm grateful for his dedication and his support these past few years of my budding career. I would also like to thank Yun as well as Professor Nir Yosef and Professor Haiyan Huang for serving on my dissertation committee. I additionally thank them for serving on my qualification exam committee, which includes Professor Satish Rao, who I had the privilege of GSI-ing for and has given me a great deal of support and encouragement. Another important mentor-figure I would like to thank is Shishi Luo, who was incredibly patient with me and walked me through the details of how to frame a project from the ground up, write and revise papers from scratch, and respond to adamant reviewers. I was truly fortunate to have been able to work closely with someone as sharp and articulate as Shishi. I would also like to thank my collaborators Matthew Mitchell and Professor Dustin Brisson for their work and guidance on selective primer design.

I am grateful to members of Professor Yosef's lab, Jim Kaminski and Shaked Afik, for their helpful feedback on my research. I have also had the pleasure of knowing past and present Song lab members Anand Bhaskar, Sara Sheehan, Ethan Jewett, Jonathan Terhorst, Zvi Rosen, Khanh Dao Duc, Jonathan Fischer, Geno Guerra, Jeff Chan, Neil Thomas, Sanjit Singh, Alan Aw, and Yutong Wang, and I would like to thank them for their helpful discussions, honest feedback, and an all-around great time. Special thanks to Jeffrey Spence, who is extremely knowledgeable and always so helpful even when he's hecticly busy. Dan Daniel Erdmann-Pham is another wildly talented individual and has been an invaluable friend to me. I'm thankful for the advice, patience, and support he has given me, even when I was certain of the contrary.

Being part of this amazing graduate program gave me the opportunity to meet a number of talented and interesting individuals. I would like to thank the faculty and administration of the EECS department, and in particular Shirley Salanio, for all the advice and consideration she has given me. A fellow graduate student I have had the privilege of knowing during my time at Berkeley is Xiang Cheng, who has always been so supportive since the day I met him in my first semester. While I haven't had the fortune of knowing Raaz Dwivedi for quite as long, I'm glad to have met him while GSI-ing, and I'm indebted to him for all the encouragement, endurance, kindness, and laughs he's given and continues to give me.

Lastly, I thank my wonderful family. I appreciate my sister and Bear for all the fun these past two years as my housemates and for cheering me up with yummy food or cute animal pictures when I'm disappointed or sad. Of course, I'd like to thank my parents who have always put my needs before theirs and worked so hard on my behalf. They have done everything they could to ensure I had a wonderful education, but most importantly, that I have a happy life.

Chapter 1

Introduction

In the past few decades, there has been a fortuitous concurrent increase in both computational processing power as well as in the volume of genomic data—the former made possible by innovation in hardware and the latter made possible by high-throughput sequencing technologies developed by companies such as Illumina, Life Technologies, Roche, and many others. Naturally, this growth has paved the way for new fields like computational genomics—the use of computational and statistical analysis to understand biology from DNA and related data. In this dissertation, we discuss two important problems in the domain of computational genomics: characterizing genetic variation in a highly repetitive region and optimizing primer sets for selective amplification.

Both projects to be discussed focus on developing computational tools using data from whole-genome sequencing (WGS). WGS refers to the sequencing of the entire genome of an organism, which is composed of A, C, G, and T nucleotides. Contrary to what its name may imply, the output of WGS is not a single contiguous sequence of nucleotides, but rather, a collection of short contiguous sequences (see Figure 1.1 for an illustration), each of which is referred to as a *read*. The length of a read is termed *read length* while the term *read coverage* or *coverage* refers to the number of reads covering a location of the genome, often meaning the average coverage across the genome. Assembling the reads into a contiguous whole genome sequence can be done by leveraging the overlap between reads and by mapping the reads to an existing whole genome (if available).

In the remainder of this chapter, we provide background information and summarize the two different problem settings to be discussed in the coming chapters.

1.1 B-cell and T-cell receptors

The first problem setting entails the characterization of genetic variation in two highly repetitive regions instrumental in the function of B-cell and T-cell receptors. B cells and T cells are both key players in the adaptive immune system and have receptors capable of initiating action or inaction when bound. For fields such as personalized medicine and evolutionary

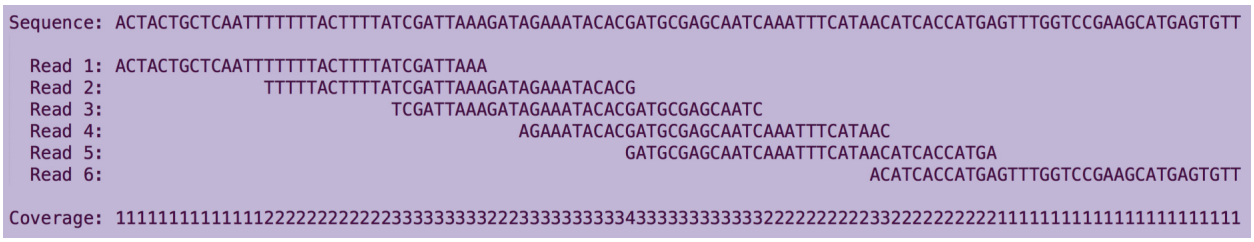


Figure 1.1: Example of sequencing data. The original sequence is shown at the top and reads 1-6 are the sequencing results. The coverage is shown at the bottom and reflects the number of reads overlapping each corresponding position.

biology, quantifying the diversity of B-cell and T-cell receptors in an individual and in the population would be incredibly valuable, but this diversity is estimated to be north of 10^{12} [2, 73]. Alternatively, instead of quantifying variation at the protein level (i.e., variation in B-cell and T-cell receptors), we can quantify the genetic diversity that contributes to this abundant receptor diversity. In this work, we focus on two regions which are both involved in the process of binding to an antigen or other molecules: the immunoglobulin heavy chain variable (IGHV) region and the T-cell receptor beta-chain variable (TRBV) region (see Figure 1.2).

IGHV and TRBV loci

The portion in the DNA which encodes for the IGHV and TRBV regions are the IGHV and TRBV loci which consist of gene families (collections of genes) which likely formed through a process of gene duplication and deletion events. While the 1 MB IGHV locus is located on chromosome 14 [71, 120], the 500 kb locus is located on chromosome 7 [96]. There are roughly 45 functional IGHV and TRBV functional genes each, where all genes are approximately 300 base pairs in length. Variation in these genes critically contributes to receptors, but to date is not carefully quantified. In this dissertation, we explore genetic variation in the IGHV and TRBV functional genes among individuals. Quantifying variation at the genomic level can have its challenges, because often the data comes post-V(D)J recombination, a process to be discussed and which may result in many missing IGHV and TRBV genes. B cells also can undergo a mechanism called somatic hypermutation, which introduces non-inherited mutations in maturing B cells in an effort to diversify and adapt receptors to new foreign elements, making the identification of the originally inherited alleles all the more difficult.

V(D)J Recombination

The process from immunoglobulin heavy chain variable (IGHV) genes to B-cell and from T-cell beta receptor variable (TRBV) genes to T-cell receptors involves a unique and defining feature of the adaptive immune system: V(D)J recombination. This series of steps is depicted

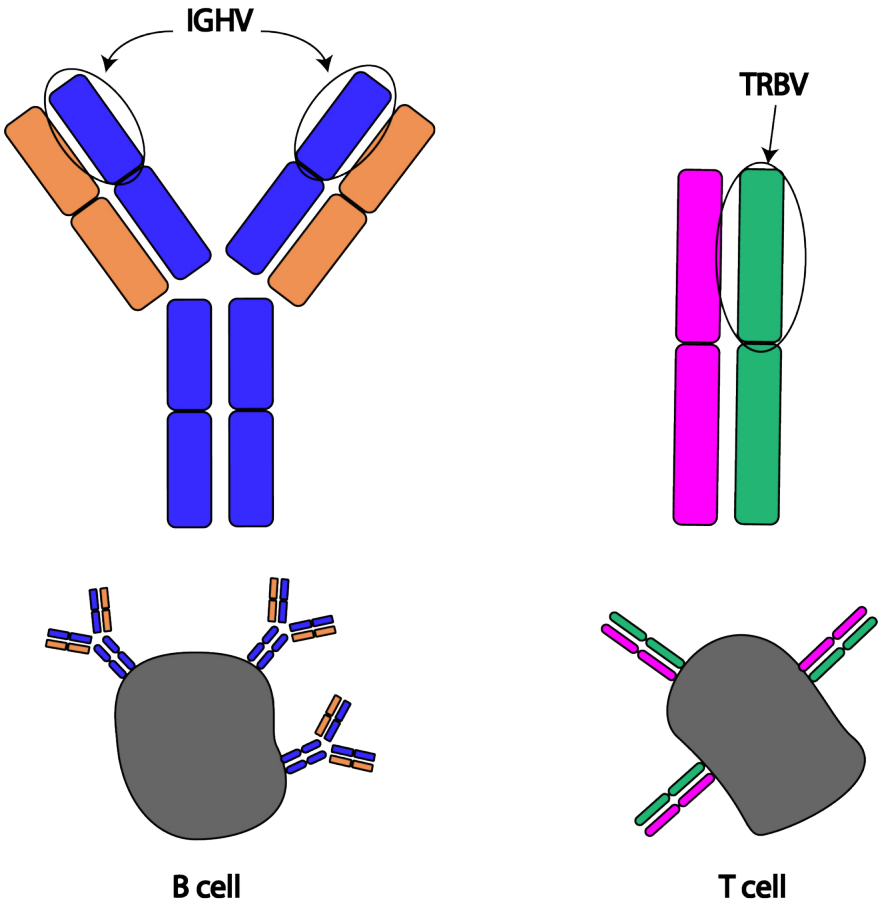


Figure 1.2: A high level schematic of the B-cell and T-cell receptor. Labels demarcate the location of the immunoglobulin heavy chain variable (IGHV) region on the B-cell receptor and the T-cell receptor beta-chain variable (TRBV) region on the T-cell receptor.

in Figure 1.3 and begins with the variable, diversity, joining, and constant genes of the immunoglobulin heavy chain locus. In the next step of the process, D-J recombination, one D gene and one J gene are stochastically selected and all genes in between are excised. Following this is V-DJ recombination, where one V gene is again stochastically selected and everything between it and the D-J combination is removed. This DNA sequence undergoes a complex process of transcription to produce RNA and translation to produce a protein product that contributes to the variable portions of the immunoglobulin heavy chain and the T-cell receptor beta-chain. This process is pivotal in contributing to the immense diversity of B-cell and T-cell receptors but creates challenges to inferring the genetic diversity of the IGHV and TRBV regions.

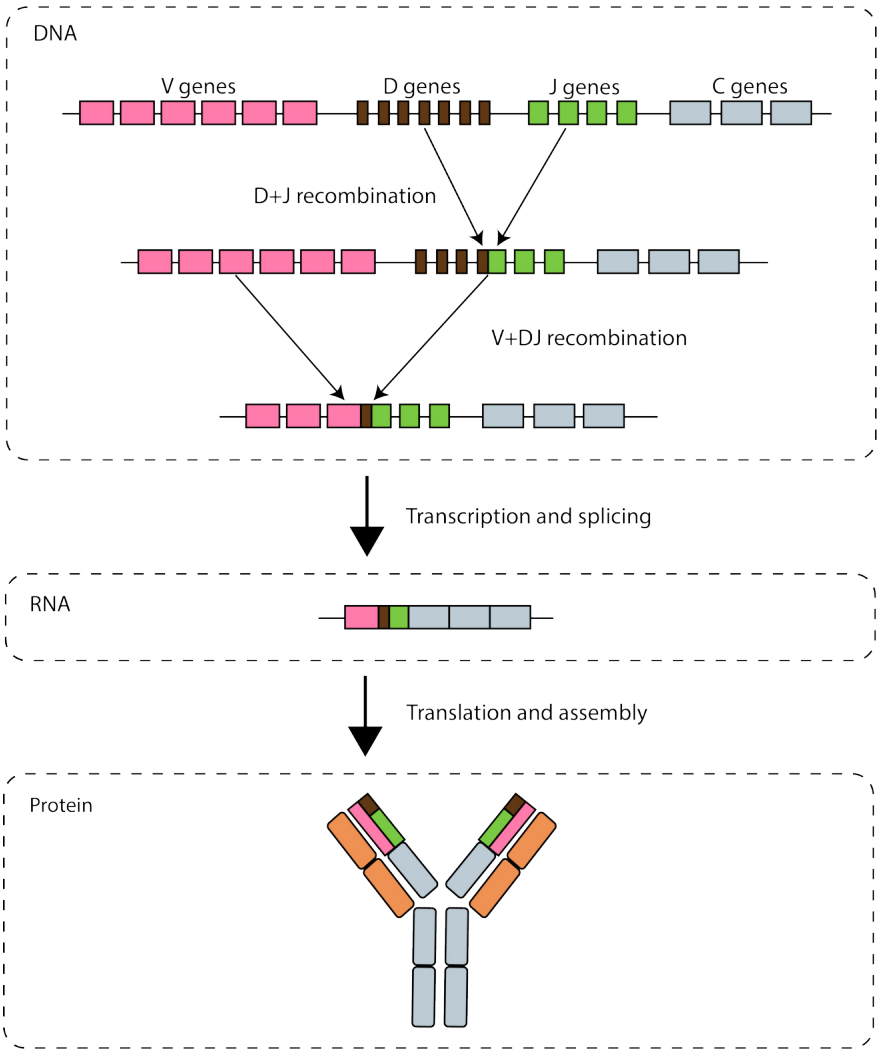


Figure 1.3: Simple illustration of the process from IGHV genes to receptor. In the first two steps, V(D)J recombination occurs, where a stochastically chosen V, D, and J gene are chosen. Everything in between the chosen V and D gene and the D and J gene is excised. Transcription and translation encode a protein product that becomes the variable region of the immunoglobulin heavy chain. Assembly with the other portions of the receptor produces the full B-cell receptor. These steps are highly simplified and serve as an introductory overview in understanding how the inherited genes contribute to the combinatorial diversity of B-cell receptors. An analogous process happens for T-cell receptors. The orange portions of the B-cell receptor are the immunoglobulin light chain, whereas the complementary portion is the immunoglobulin heavy chain.

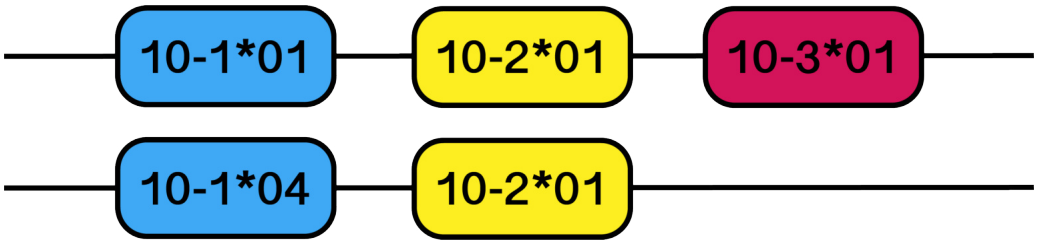


Figure 1.4: An example of two haplotypes from the same organism with genes *10-1*, *10-2*, and *10-3* and the * indicates the allele in this dissertation.

Genomics definitions

Lastly, we clarify our usage of three terms commonly used in biology.

- *gene*: A subsequence of the genome which encodes instructions for some function.
- *allele*: A specific variant of a gene (indicated by a *). For example, *1-69*01* indicates allele '01' of the gene *1-69*.
- *haplotype*: A collection of alleles belonging to the one chromosome copy. Humans, for example, have two copies of each chromosome.
- *somatic mutations*: mutations in an organism that are not passed on to its offspring. Somatic hypermutation, for example, introduces mutations which are not passed on to offspring but is an important process of the adaptive immune system.
- *germline mutations*: mutations in an organism that are passed on to offspring.
- *genotype*: identification of the genetic constitution of an individual organism.

For an illustrative example of the first three definitions, see Figure 1.4 which depicts two haplotypes. One haplotype has all three genes *10-1*, *10-2*, and *10-3* whereas the other haplotype is missing gene *10-3*. Additionally, the top haplotype has allele *10-1*01* whereas the other haplotype has *10-1*02*, but for the gene *10-2*, the two haplotypes have the same allele. In addition to deletion events, duplication events may also occur, meaning a gene or set of genes is duplicated on the same haplotype.

This concludes background information for Chapters 2 and 3 of this dissertation. For a more thorough background, [76] is a great resource. We now discuss background information for Chapter 4.

1.2 Selective Whole Genome Amplification (SWGA)

This next problem setting is motivated by the lack of methods needed to sequence a specific genome of interest. For example, it is often the case that a particular microbial species of interest is cultivated from a complex, natural sample. This particular species may comprise less than 1% of the entire sample, making it very inefficient to sequence. In many cases, in vivo cultivation of the species can be done via a host organism. For *Wolbachia*, an endosymbiotic bacteria compatible with the fruit fly, one would need roughly 1000 live adult flies to cultivate enough *Wolbachia* to achieve roughly 10x coverage. This method can be extremely cumbersome and can also sometimes be legally unethical, as is the case with *Plasmodium reichenowi*, which exists in chimpanzees infected with malaria. Often times the target genome cannot even be cultured in the lab, necessitating other methods for amplifying the target genome.

Multiple Displacement Amplification

In 2001, Dean et al [25] proposed a method of using the enzyme ϕ29 DNA polymerase and primers (short DNA sequences, composed of A, G, T, C nucleotides) to amplify whole genomes via multiple displacement amplification. As depicted in Figure 1.5, this process starts with the primers binding to the target genome. In the subsequent steps, the ϕ29 proceeds along the genome, synthesizing the complementary strand. When synthesis is impeded from encountering a primer bound to the genome, the ϕ29 enzyme will displace the synthesized strand and continue its own synthesis. The same process can be repeated on these "branched" strands, creating the opportunity for exponential amplification. This procedure using multiple displacement amplification supplants the need for lengthy growth periods and traditional DNA isolation methods. However, this solution has its own challenges. Namely, typical genomes can be thousands to billions of base pairs long, mandating computational tools for processing these large genomes and for designing sets of primers that will optimally amplify the target genome.

1.3 Summary

Now that we have established the necessary background information, we outline the chapters of this dissertation. In Chapter 2, we discuss the development of our model for estimating copy number and allelic variation in the IGHV locus. Lack of standard methods to genotype this region prevents it from being included in association studies and is impeding the growing field of antibody repertoire analysis. In general, the study of genomic regions that contain gene copies and structural variation is a major challenge in modern genomics and unlike variation involving single nucleotide changes, data on the variation of copy number is difficult to collect. Our method presented in Chapter 2 takes short reads from high-throughput sequencing and outputs a genetic profile of the IGHV locus with the read coverage depth and

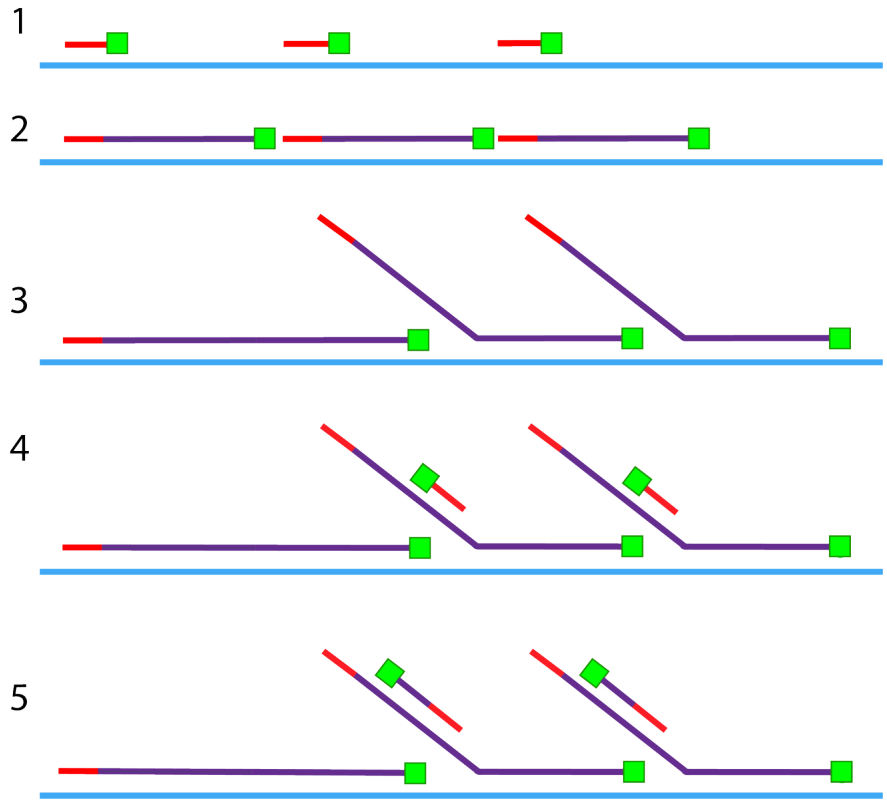


Figure 1.5: Illustration of multiple displacement amplification. Red lines correspond to primers, green to ϕ29 enzymes, blue to the genome to be amplified, and purple to the the polymerized DNA. In the first and second steps, the primer binds to regions in the genome and with the help of enzyme ϕ29, it polymerizes a complementary strand. This continues and displaces downstream strands, allowing for additional amplification on branches, as shown in steps 4 and 5.

a putative nucleotide sequence for each operationally defined gene cluster, without the need to reconstruct the complete sequence for the region. Tests on simulated data demonstrate that our approach can accurately determine the presence or absence of a gene cluster from reads as short as 70 bp. When applied to a family composing three generations, our pipeline outputs genotypes that are consistent with the family pedigree, confirms existing multigene variants, and suggests new copy number variants. This study paves the way for analyzing population-level patterns of variation in IGHV gene clusters in larger diverse datasets and for quantitatively handling regions of copy number variation in other structurally varying and complex loci.

In Chapter 3, we build upon the model developed for IGHV in Chapter 2 by extending our analysis to the T cell beta variable (TRBV) locus, another complex and variable region in the

human genome, and by refining our detection of single nucleotide polymorphisms and novel alleles. We also 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 from over a 100 populations. From our estimates of copy number, allelic, and single nucleotide variant frequencies across geographic regions in a sample of unprecedented size for the IGHV and TRBV loci, we find that the IGHV and TRBV gene families exhibit starkly different patterns of variation—namely, that there is strong evidence the IGHV locus undergoes more frequent gene duplication and deletion than the TRBV locus but that the TRBV locus may exhibit higher rates of nucleotide substitution. Support from data in other vertebrate species also indicates that the evolutionary dynamics we infer from the IGHV and TRBV loci in humans were also in force over longer, macro-evolutionary, time periods. Our work provides a quantitative analysis of genetic variation relevant to the fields of population genetics and, medicine, and a broad spectrum of scientific research as well as being of importance to understanding the long-term evolution of these gene families.

In Chapter 4 we tackle the problem of selective whole genome amplification. The cost of genomic sequencing over time has seen enormous decreases in the past two decades, but the process of sequencing complex DNA samples still remains sub-optimal. In particular, in settings where the interested genome is difficult to isolate and comprises a minuscule fraction of a heterogeneous DNA sample, the sequencing effort will be vastly disproportionate to the amount of sequencing of the targeted genome, wasting both time and resources. In this chapter, we describe SOAPSwga, an optimized and parallelized pipeline for primer design in the SWGA context. Unlike previous methods, SOAPSwga incorporates appropriate machine learning and active learning methods to model primer efficacy using thermodynamically-principled calculations of binding affinities. Additionally, when evaluating and searching for primer sets, SOAPSwga explores according to branch-and-bound principles to pursue only the most promising sets via a data-driven evaluation model incorporating novel features rather than performing brute force search according to a human-inferred heuristic function. In this paper, we discuss this pipeline and our design of primers sets for the selective amplification of *Mycobacterium tuberculosis* in a sample of human blood. Lastly, we expand on the importance of this work to many fields in genomics, and in general, its potential usefulness to any setting consisting of targeted sequencing.

Chapter 2

Estimating variation at the IGHV locus

This chapter is joint work with Shishi Luo and Yun S. Song and appears in *PLoS Computational Biology* [64].

2.1 Introduction

The variation between human genomes in gene copy number is understudied and poorly characterized. One such region where this variation is known to exist is the immunoglobulin heavy variable (IGHV) locus. It is a vital component of the adaptive immune system, containing the V genes that code for a component of the heavy chain of antibody molecules. Like other multigene receptor families, the gene segments in this region have been accumulated over time through a process of gene duplication and diversification [83, 80, 24]. As such, many of the genes in this locus are highly similar and there are repetitive DNA elements interspersed throughout the region. IGHV haplotypes (instances of the IGHV locus) vary not only by single nucleotide polymorphisms but also in the copy number and ordering of gene segments [121]. All these characteristics make it difficult to determine the nucleotide sequence of this region and, to date, only two full sequences of the IGHV locus exist [71, 121]. Despite the increasing affordability of whole-genome sequencing (WGS), there are currently no methods to genotype the IGHV locus directly from WGS reads (for a tool that extracts genotypes from long contigs, see [82]). Thus, basic population-level characteristics of the locus, such as the mean number or standard deviation of copies of gene segments have not yet been quantified.

The human IGHV locus lies at the telomeric end of chromosome 14 and is approximately 1 Mb in length. In this 1 Mb region, there are about 45 functional genes, each approximately 300 bp in length. There are also approximately 80 non-functional pseudogenes in the region, so-called because they are either truncated or contain premature stop codons. Known allelic variants of individual IGHV genes are currently curated in the International Immunogenetics

Information System (IMGT) Repertoire database [37]. The standard nomenclature for IGHV genes is detailed in Section 2.3.

Given the role of the IGHV locus in the adaptive immune response, IGHV haplotypes are obvious candidates as genetic determinants for susceptibility to infectious disease. Several early targeted studies of the IGHV locus have implicated allelic variation and copy number in determining expressed antibodies repertoires and understanding disease susceptibility [75, 21, 16, 88, 48, 102, 120]. Allele *3-23∗03*, for example, has been shown to be more effective in binding Haemophilus influenzae type (Hib) polysaccharide than the most common allele, *3-23∗01* [61]. Despite such findings, however, the IGHV locus is rarely included in genome-wide association studies, largely in part to the lack of standard format and tools to quantitatively characterize variation in the region.

Lack of tools for genotyping the IGHV locus also hampers the burgeoning field of antibody repertoire sequencing [52, 95, 35, 12, 122], which is being used in numerous medical applications, including inferring the evolutionary path of broad and potent monoclonal antibodies against human immunodeficiency virus (HIV) [123, 60, 27], detecting blood cancers [94, 32], assessing the impact of aging on the antibody response [43], and measuring the adaptive immune response to vaccination [44, 41]. The first step in many of these studies is to align each read, sequenced from the antibody repertoire of an individual to its germline gene. The current practice is to use germline alleles in a public database of all known alleles (such as the IMGT Repertoire database) for alignment. This is a severe limitation of the process because after undergoing somatic hypermutation, antibody sequences may be so different from the germline that the top-matching allele in the database no longer corresponds to the germline allele in the individual.

Here, we address the pressing need for methods to genotype the IGHV locus. By leveraging the IMGT database of known alleles and the increasing availability of WGS data, we construct a pipeline that determines the functional genes present in an individual's IGHV locus from short reads. Not only is our approach high-throughput, but it also produces output that is annotated and in a format ready for quantitative comparisons between multiple individuals. With reads as short as 70bp and with coverage of 30x, our pipeline accurately detects the presence of gene segments from simulated reads of the two known IGHV references (GRCh37 and GRCh38). With sufficiently long read lengths (250bp), our pipeline also outputs accurate nucleotide sequences of gene segments present in single copy. We then run our pipeline on an empirical dataset of whole-genome sequencing reads from a sixteen member family, obtaining for the first time distributions of copy numbers in this family. Our copy number calls are consistent with the family pedigree and confirm known multigene variants of the IGHV locus. Our results also suggest evidence of new haplotypes that are mosaics of the existing reference haplotypes and haplotypes that might be transitional between them.

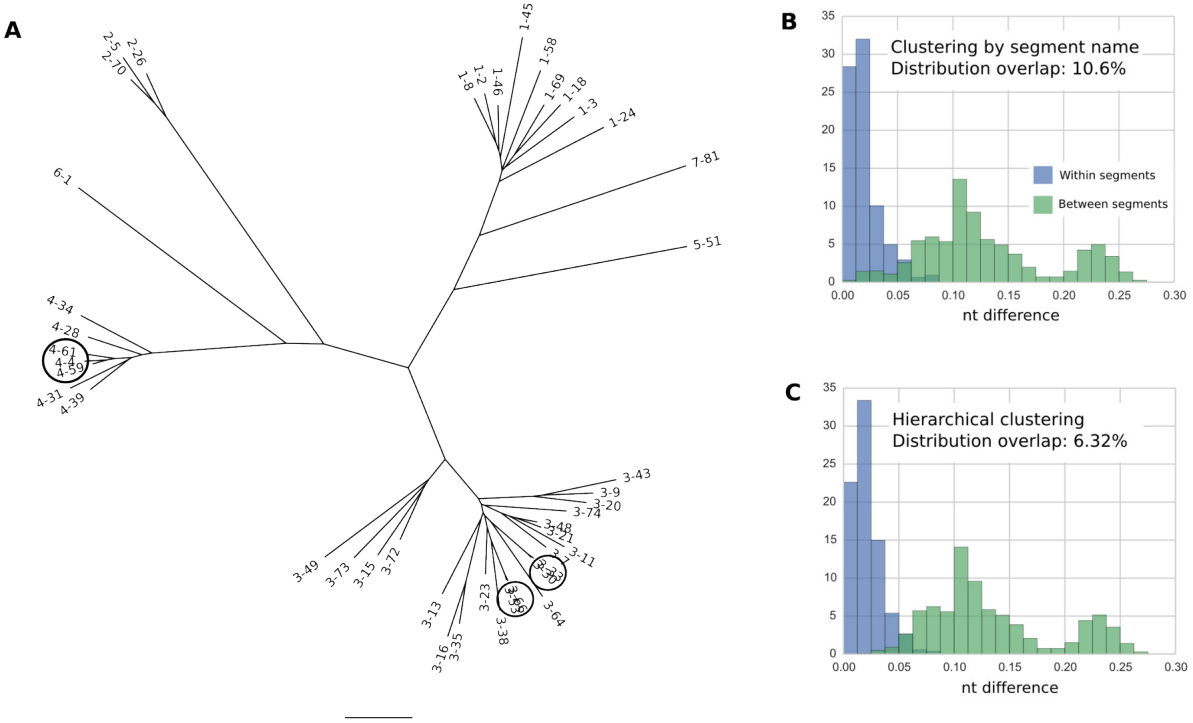


Figure 2.1: Alleles clustered according to nucleotide similarity. (A) Phylogenetic tree reconstruction of the gene segments in the haplotype sequenced in [71]. Circles highlight alleles that are evolutionarily very close. Tree made using neighbor-joining method in `ape` package in R based on Hamming distance between multiple sequence alignment. (Phylogenetic reconstruction using BEAST [28] led to a qualitatively similar tree). Allele numbers are suppressed for clarity. (B) Distribution of percent nucleotide difference (Hamming distance divided by alignment length) between alleles from same IMGT segment (blue) compared against alleles from different segments (green). Alleles from duplicate segments (e.g. *1-69* and *1-69D*) have been merged for this analysis. (C) Same as (B) but with alleles partitioned by results of hierarchical clustering rather than IMGT segment name.

2.2 Results

Hierarchical clustering to define operational segments

The main difficulty in accurately genotyping the IGHV locus is the high level of similarity between alleles of different gene segments. Figure 2.1A illustrates the level of nucleotide similarity between the IGHV segments in GRCh37. For example, the alleles of segments *3-30* and *3-33* in GRCh37, circled in Figure 2.1A, differ in only 1.4% of their nucleotides. Since some segments have alleles that differ by more than 1.4% in their base pairs (Figure 2.1B), it

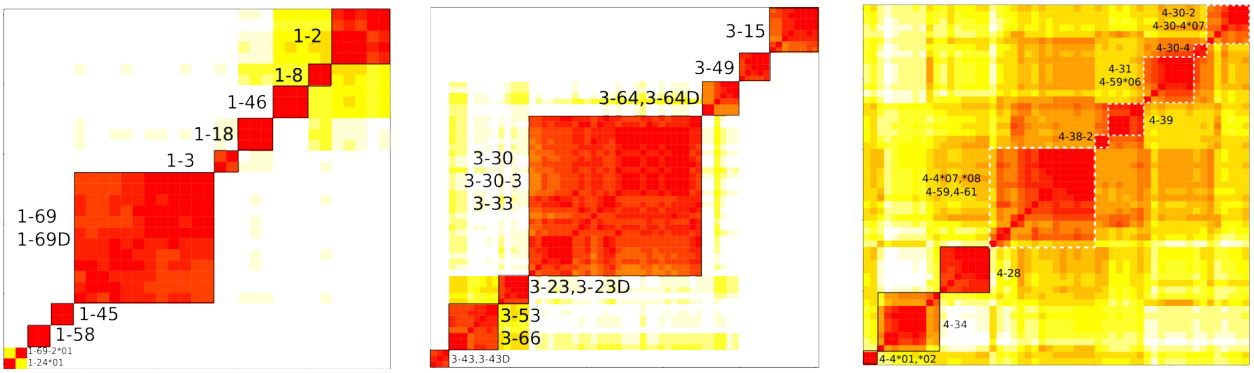


Figure 2.2: Heatmaps of matrices of Hamming distance between alleles. Rows and columns are ordered according to clusters found by hierarchical clustering as described in Section 2.3. Color spectrum ranges linearly from red to white for nucleotide distances 0-10%. Differences greater than 10% are white. (Left) Alleles from family 1. (Center) Alleles from family 3. Full set of alleles in Figure A.3. (Right) Alleles in family 4. Dashed white squares indicate possible clusters.

becomes problematic to distinguish between alleles of the GRCh37 genes *3-30* and *3-33* based on nucleotide dissimilarity. To be more concrete, if one had reads of length 100 bp from a haplotype containing both *3-30* and *3-33* segments, it would be algorithmically very difficult, if not impossible, to correctly map reads that are from regions common to 3-30 and *3-33*. We note that this difficulty in distinguishing between alleles becomes even more pronounced when analyzing antibody repertoire sequencing data, where somatic hypermutation further confounds the matching of repertoire sequences to germline alleles [126].

This problem also occurs with other gene segments: across all full-length functional IMGT alleles, there is a 10.6% overlap in the distribution of nucleotide differences between alleles with the same segment name and alleles with distinct segment names (Figure 2.1B). Reads from the alleles in this overlapping region cannot be operationally distinguished from each other, leading to unreliable and ambiguous genotype calls. Thus, it does not make sense to keep these alleles separate and we pool them together into units we call "operational segments". As we show in the next sections, this strategy allows us to extract useful information, such as copy number estimates, with less ambiguity.

To determine these operational segments in a systematic manner, we perform hierarchical clustering within each family of full-length, functional IMGT alleles (Section 2.3). This groups the alleles together according to their sequence similarity. As a result of this clustering, we reduce the overlap compared to clustering by segment name alone (Figure 2.1C). From this figure, we see that although we cannot eliminate the overlap completely, in most operational segments, alleles are within 5% nucleotide differences of each other.

Some families have clearly defined operational segments. In family 1, the clusters correspond to segment name, as long as duplicate segments such as *1-69D* and *1-69* are merged

(Figure 2.2). In families 2 and 5, which have three and two segments respectively, the alleles cluster by segment name (Figure A.1, Figure A.2). In family 3, five segments that have distinct names—namely, *3-30*, *3-30-3*, *3-33*, *3-53*, and *3-66*—form two clusters {*3-30*, *3-30-3*, *3-33*} and {*3-53*, *3-66*} (Figure 2.2). Families 6 and 7 each have only one segment and therefore do not require clustering.

Surprisingly, the same clustering algorithm that leads to clean clusters in the other families fails to identify clear-cut clusters in family 4 (Figure 2.2). This is the main source of overlap still seen in Figure 2.1C. Not only are the boundaries between clusters fuzzy in this case, but alleles of the same segment cluster separately. For example, *4-4*01* and *4-4*02* cluster separately from *4-4*07* and *4-4*08*. The alleles in family 4 also seem to be more similar to each other than alleles in other families. It is not clear why alleles in family 4, in particular, should cluster poorly compared to the other families. Gene conversion events in IGHV family 4 and a more recent common ancestor than other IGHV families are both possible explanations that are consistent with the observed distance matrix. A better clustering, based on a combination of mutational distance and indel distance, was ultimately used to define the operational segments for family 4 (Figure A.4).

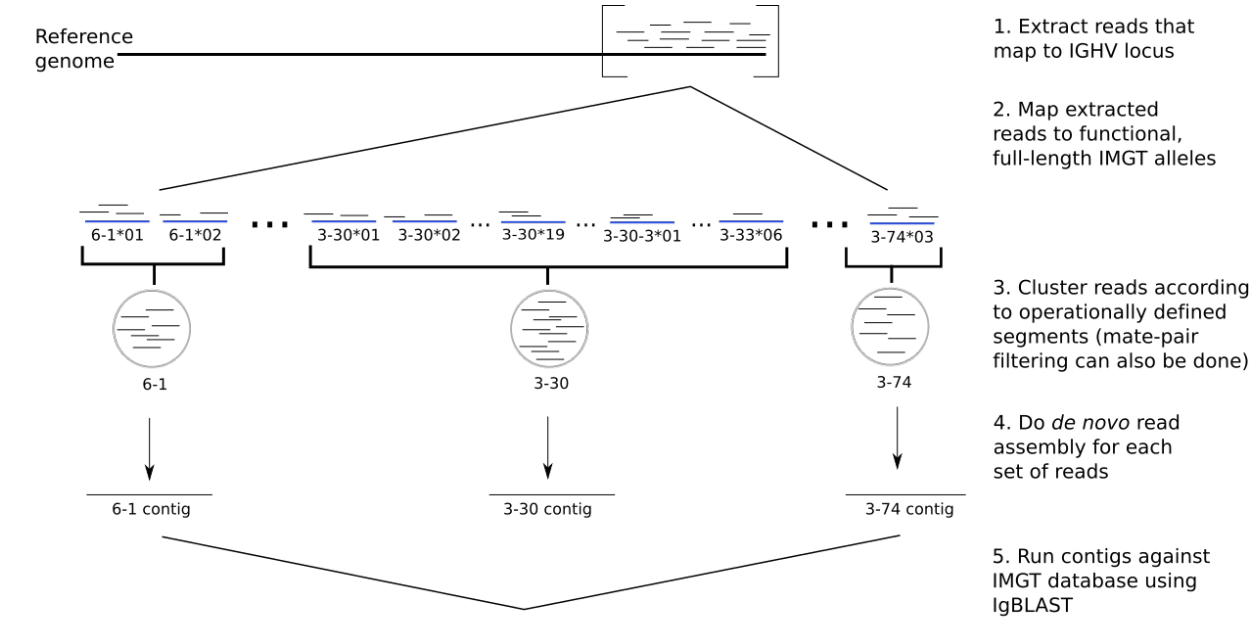
With the caveat that family 4 gene segments are more speculative, Table 2.1 summarizes the operational segments as defined by hierarchical clustering. Only segments for which the alleles differ from the current IMGT nomenclature are listed. For the remainder of this article, unless stated otherwise, we will use the segment names as they are defined in Table 2.1.

Pipeline performance on simulated reads

The operational gene segments (Table 2.1) address the main difficulty in genotyping the IGHV locus and is the key idea behind our data pipeline (Figure 2.3). Without this crucial step, it is difficult to determine IGHV alleles from read mapping alone (Table A.1). The input of the pipeline is a file of whole-genome sequencing reads from an individual and the output consists of a segment-by-segment reconstruction of the IGHV locus, with a point estimate of copy number, the closest matching existing IMGT allele, and a nucleotide sequence reconstruction of each operational segment. Figure 2.4 shows the performance of our pipeline at three levels of genotype resolution on simulated reads from the two complete IGHV haplotype sequences (Section 2.3).

At the coarsest scale, we ask whether the pipeline correctly identifies the presence or absence of each operational segment. We find the pipeline to be highly accurate, with a precision of 100% for all coverage depth (30×, 40×, 50×) and read length (70 bp, 100 bp, 250 bp) combinations. This means that all the segments identified by our pipeline are in the reference haplotype. The recall, the fraction of segments in the reference that are identified by our pipeline, is 100% for all but two of the coverage depth/read length combinations (Figure 2.4A).

At the next level of resolution, we ask whether the pipeline can correctly determine the copy number of each operational segment. We use the read coverage depth of the assembled



seg	top_match	length	muts	indels	cov	seq
1-18	1-18*01	296	0	0	5.4186	CCCACTCCCAGGTTCAGCTGGTGCAGTCTGGAGCTGAGGTGAAGAA...
1-2	1-2*02	261	1	0	5.99257	CTGGGGCTGAGGTGAAGAAGCCTGGGGCCTCAGTGAAGGTCTCCTG...
1-24	1-24*01	296	2	0	16.8413	CCTCTCCACAGGCACCCACGCCCAGGTCCAGCTGGTACAGTCTGGG...
1-45	1-45*02	296	0	0	29.8786	CCTCCACAGATGCCTACTCCCAGATGCAGCTGGTGCAGTCTGGGGC...
1-46	1-46*01	296	0	0	30.8161	CCTCTCCGCAGGTGCTCACTCCCAGGTGCAGCTGGTGCAGTCTGGG...
1-58	1-58*01	296	0	0	33.6077	CCACAGGTGCCCACTCCCAAATGCAGCTGGTGCAGTCTGGGCCTGA...
1-69	1-69*01	296	0	0	31.4459	CCTCCACAGGTGTCCAGTCCCAGGTGCAGCTGGTGCAGTCTGGGGC...
2-26	2-26*01	301	0	0	18.7716	CTTCTCCACAGGGGTCTTGTCCCAGGTCACCTTGAAGGAGTCTGGT...

Figure 2.3: Schematic of the genotyping pipeline. 1) WGS reads (short thin black horizontal lines) that map to the IGHV locus of a single individual are extracted from the full set of reads. 2) These extracted reads are mapped to known functional IGHV gene segment alleles (thick blue horizontal lines) curated from the IMGT database. 3) Mapped reads are pooled according to the operational segments described in the Results section. At this stage, extra filtering, for example using mate-pair data, can also be applied. 4) Local assembly is performed on reads to produce contigs (long thin black horizontal lines) corresponding to each operational segment. 5) The resulting contigs are identified using stand-alone IgBLAST [124]. The final output contains, for each individual and each assembled contig: the closest-matching existing allele, the length of the match, the number of nucleotide mutations or indels that separate the contig from the closest-matching allele, the read coverage of the contig as reported by SPAdes, and the nucleotide sequence of the contig.

A

GRCh37	70bp	100bp	250bp
30x	1.0	1.0	0.97
40x	1.0	1.0	1.0
50x	1.0	1.0	1.0
GRCh38	70bp	100bp	250bp
30x	1.0	1.0	1.0
40x	1.0	1.0	1.0
50x	0.97	1.0	1.0

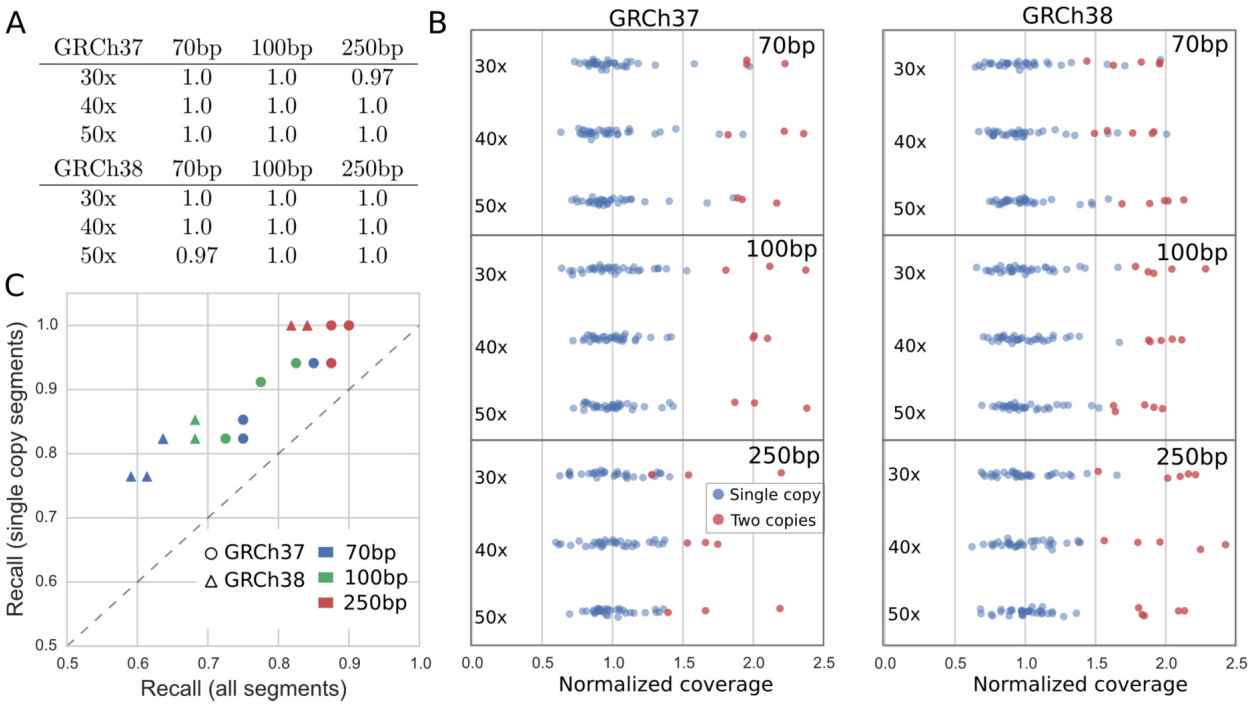


Figure 2.4: Performance of pipeline on simulated reads from GRCh37 and GRCh38 for varying coverage depths and read lengths. (A) Recall fractions of the pipeline for the two human reference genomes (precision fractions are all 1 and not shown). Recall is calculated as the fraction of operational segments in the reference genome that are correctly called by the pipeline. (B) Read coverage depth of each assembled segment (the point estimate for copy number) colored by actual copy number (detailed in Table A.2) in the reference genome. Raw read coverage depth has been normalized by the average coverage of single-copy segments. Jitter has been applied to the vertical coordinates to better show their distribution. (C) The recall of alleles for all segments versus the recall for single-copy segments. Each point is one reference genome, coverage depth, read length combination. Note that two red triangles overlap at the point $(0.84, 1.0)$. Different coverage depths are not indicated because there is no pattern between coverage depth and allele reconstruction accuracy.

Table 2.1: Operational gene segments as defined by our hierarchical clustering analysis. Only those that differ from the current IMGT nomenclature are listed, i.e. alleles that cluster by their IMGT segment name are not shown.

Operational name	Alleles, under current IMGT nomenclature
1-69	All 1-69 and 1-69D alleles
2-70	All 2-70 and 2-70D alleles
3-23	All 3-23 and 3-23D alleles
3-30	All 3-30, 3-30-3, and 3-33 alleles
3-43	All 3-43 and 3-43D alleles
3-53	All 3-53 and 3-66 alleles
3-64	All 3-64 and 3-64D alleles
4-4∗01	4-4∗01, 4-4∗02
4-30-2	All 4-30-2 alleles and 4-30-4∗07
4-31	4-30-4∗01, 4-30-4∗02, 4-31∗01-∗04, 4-31∗10
4-31∗05	4-31∗05
4-59	4-4∗07, 4-4∗08, and all 4-59 alleles
4-61	4-61∗01, 4-61∗03-∗05, 4-61∗08
4-61∗02	4-61∗02

contig as our point estimate for copy number. Figure 2.4B shows that contig coverage depth is indeed correlated with copy number, though there is variation above and below the true copy number and some segments which are present in a single copy have high coverage depth. This is because pseudogenes in the IGHV locus, which are not included in our reference set, may share common subsequences with functional genes. Reads from pseudogenes can, therefore, be erroneously mapped, artificially inflating the contig coverage depth. This is particularly an issue with 70 bp length reads as these reads are more likely to completely fall within a conserved region. This problem can be partly alleviated with paired-end reads, a strategy we use on the real dataset in the next section.

At the highest level of resolution, we compare the assembled contig obtained from the pipeline to the known nucleotide sequence for each segment. When a segment is only present in single copy in the locus, and if the read lengths are 250 bp, the recall of the segment nucleotide sequence is 100% in all but one of the simulated datasets (Figure 2.4C). With shorter reads, the frequency of correctly calling alleles is lower. As with copy number determination, this lower accuracy is likely due to erroneously mapped reads from pseudogenes and highly similar functional genes that interfere with the assembly algorithm. For the same reason, when a segment is present in more than one copy and as different alleles, the allele calls are also less accurate. Note that higher coverage depth does not necessarily improve accuracy because the error arises not from sequencing error, which occurs in random locations and can be mitigated with higher coverage depth, but from erroneously mapped reads, which are

systematically incorrect regardless of coverage.

Genotyping the Platinum Genomes dataset

We next apply the pipeline to the publicly available Platinum Genomes dataset [29], a set of whole-genome sequencing reads of length 100 bp at roughly 30× coverage depth from a family of 16 individuals (four grandparents, a mother, a father, and ten children, all of European ancestry). Because these reads are paired, we perform an additional filtering step (Section 2.3) to discard reads that are potentially from pseudogenes (Figure A.12) in order to improve our allele calls and decrease the false discovery of duplicated genes.

A summary of copy number and allelic variation in IGHV segment types in this dataset is shown in Figure 2.5. For all the results that follow, the raw coverage depth of each segment is scaled by the coverage depth of segment *3-74* in the same individual to eliminate variation due to differences in read coverage between individuals. We choose segment *3-74* because it has no documented examples of copy number variation and is located at the telemeric end of the chromosome. Specifically, we assume that *3-74* has two copies, one on each chromosome, and divide the coverage depth of all other segments by half of the coverage depth of *3-74*. A normalized coverage depth of 1, therefore, corresponds to single copy on one, but not both, of the chromosomes. Note that the coverage depth tends to decrease towards the *6-1* end of the locus due to VDJ recombination, an issue we will return to in the Discussion.

General patterns of variation. Figure 2.5 shows for the first time the distribution of copy number variation in a sample of individuals, segment by segment. Variation exists in copy number within segments and between segments. Some, such as *3-72* and *3-73* are present in all individuals as two copies, one on each chromosome. Others, such as *1-8*, *3-9*, *5-10-1*, *4-38-2*, and *1-69-2*, are either absent (coverage of zero) or present as single copy on one chromosome. We note that even though the Platinum Genome reads were mapped to GRCh37, we are nevertheless able to assemble full nucleotide sequences of *4-38-2* and *1-69-2* that are not in the reference and which are quite distinct from all other alleles. Normalized coverage around the value of three or higher indicates a segment has a duplicate on the same chromosome (segments shaded in grey in Figure 2.5). These include *3-23*, *3-30*, *4-31*, *3-43*, *1-46*, *3-53*, *3-64*, *1-69*, and *2-70*, which are known to have duplicates, but also *1-24*, *4-34*, *1-45*, *3-49*, and *1-58*, for which duplicates have not been previously documented. These latter gene segments are new candidates for copy number variants and topics for further study.

Thirteen out of the forty-two segments (about 30%) found in the family are each represented by the same single allele in all sixteen members of the family. The unique alleles corresponding to these segments are denoted in Figure 2.5 along the top of the plot. This strongly suggests that these segments are homozygous in the four unrelated grandparents. Genotyping of a larger sample will ascertain whether this set of common alleles is shared for all individuals of European ancestry or is a by-product of our sample being for a small pedigree. In either case, our pipeline and approach begin to address the question of whether a subpopulation can be uniquely identified by a common set of IGHV alleles.

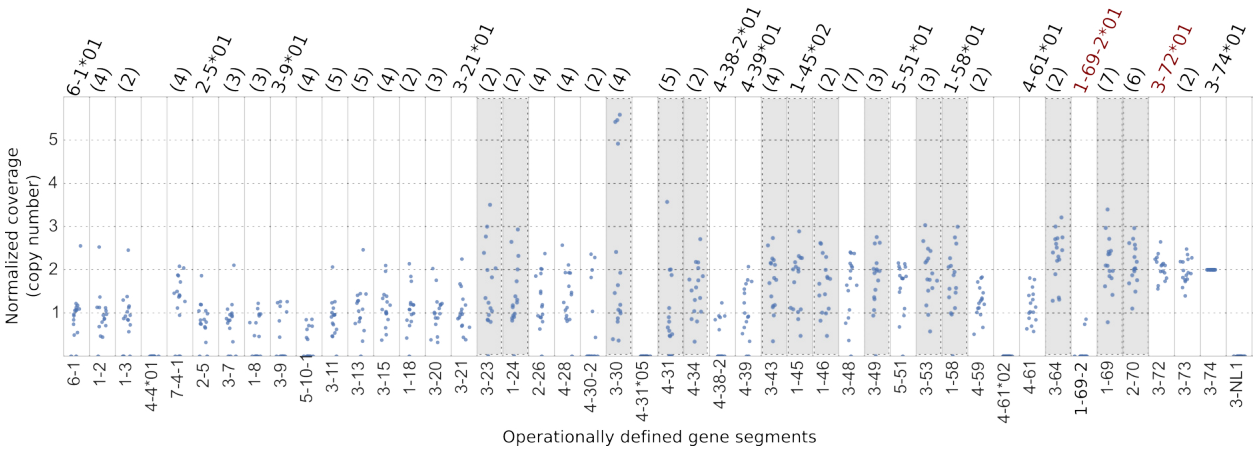


Figure 2.5: Dotplots of coverage calls for each segment type for Platinum Genomes data. The Y-axis is the normalized coverage, i.e. the coverage depth divided by half of the coverage depth of segment type *3-74* (assumed to have two copies). Segment types are ordered, where possible, according to their location in the genome, from *6-1* (centromeric end) to *3-74* (telomeric end). The number in parentheses above each segment type is the number of unique allele sequences found in the family. If only one allele was found, its name is given (in the case where a segment has only one known allele in the IMGT database, the allele name is in red). Shaded columns indicate segment types that likely have more than one copy per chromosome. The outliers for segment types *6-1*, *1-2*, and *1-3* all correspond to a single individual, NA12891, who had relatively uniform coverage over all segments (Figure 2.8, Figure A.5). Horizontal jitter has been applied to all points to better illustrate the distribution.

Multigene copy number variants. We next looked for the presence in family members of two copy number variants involving multiple gene segments that differ between the GRCh37 and GRCh38 reference haplotypes (Figure 2.6). Using knowledge of the family pedigree, we are able to reconstruct the diploid genotype for these variants.

In the case of alternative haplotypes *1-8/3-9* (GRCh37) and *3-64D/5-10-1* (GRCh38), our point estimates for copy number show that maternal grandparent NA12891 carries both configurations, one on each chromosome. In contrast, the *1-8/3-9* type is entirely absent from the paternal side of the family (Figure 2.6A). We manually checked our pipeline output to verify that the copy number calls in the children are consistent with the pedigree. Indeed, both NA12881 and NA12888, which appear to be missing the *3-9* segment, generated reads that mapped to a full-length *3-9∗01* allele (indicated by '∗' in plots). This is consistent with NA12881 carrying the GRCh37, but not GRCh38, configuration and NA12888 carrying both the GRCh37 and GRCh38 configurations. (Our automated pipeline did not call the *3-9* segment because the coverage of that segment was too low for the assembler to run). We also verified that NA12890 and NA12886 do not carry the *5-10-1* segment, suggesting

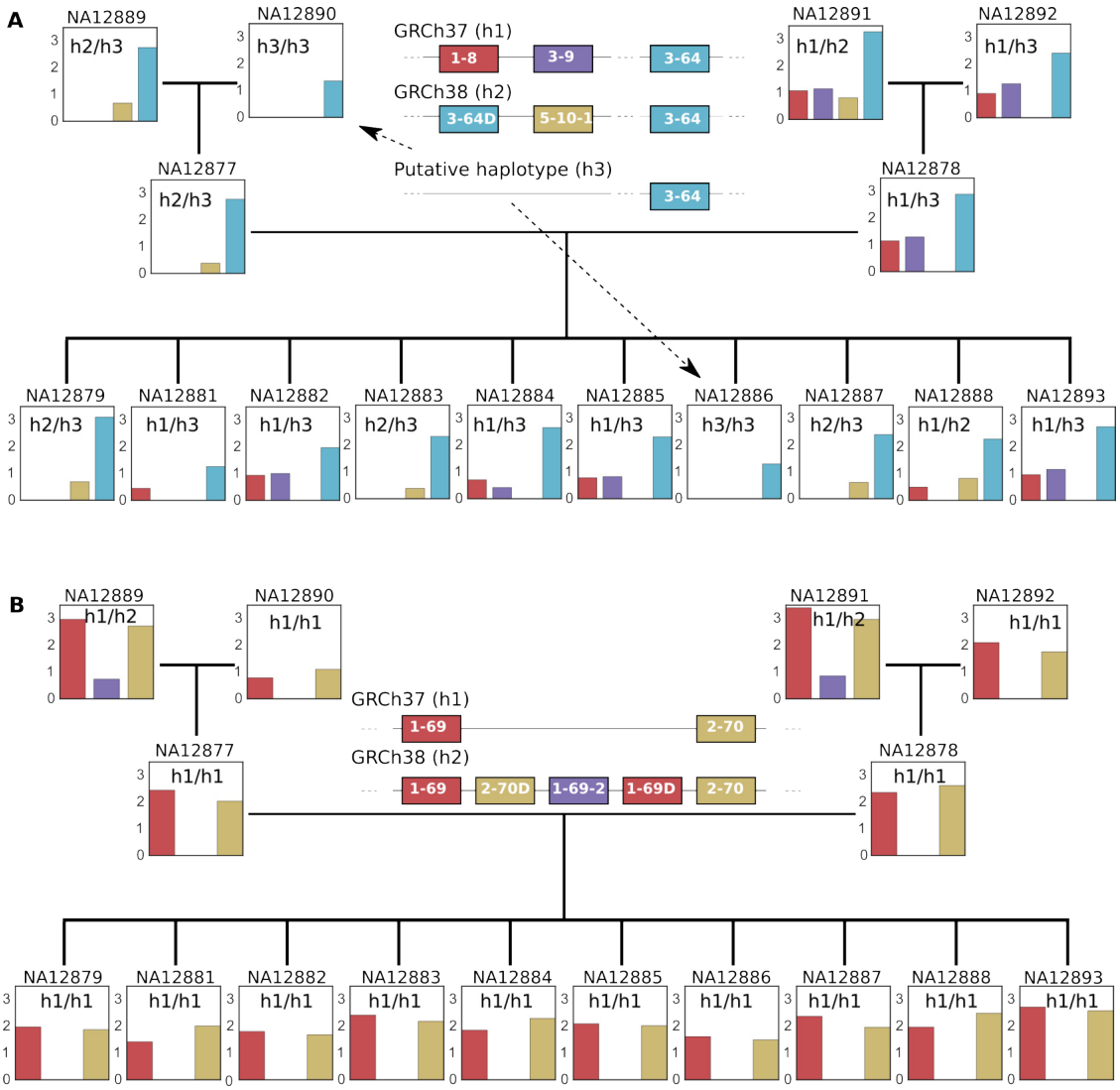


Figure 2.6: Coverage calls for two multigene copy number variants in GRCh37 and GRCh38. (A) Alternative haplotypes *1-8/3-9* and *3-64D/5-10-1*. Additional manual examination of pipeline output shows, consistent with the putative diploid status, that 3-9 is present in NA12881 and NA12888 (indicated by stars '$*$') and that 5-10-1 is not present in NA12886. (B) The insertion haplotype *2-70D/1-69-2/1-69D*. In both subfigures, Y axis on each bar plot is normalized coverage. Each bar is colored according to the segment it corresponds to. The putative diploid status for each individual is indicated on the coverage bar plot. Individuals are arranged according to their family tree. NA12877 and NA12878 are the father and mother respectively.

a new haplotype, transitional between GRCh37 and GRCh38, which contains a single *3-64* segment without either the *1-8/3-9* or *3-64D/5-10-1* gene combinations. In both these individuals, there were allele calls and positive coverage for genes towards the *6-1* end of the locus, indicating that the absence of *1-8/3-9* and *3-64D/5-10-1* genes are not due to VDJ recombination in the cell type or low coverage (Figure A.6). Further, these two individuals appear to be homozygous for this new haplotype, suggesting that the haplotype is common.

For another multigene copy number variant, the *2-70D/1-69-2/1-69D* insertion (on GRCh38 but not on GRCh37), grandparents NA12889 and NA12891 carry the insertion on one chromosome and not on the other (Figure 2.6B). The insertion did not transmit to the parents or children, with neither the presence of *1-69-2* or elevated coverage for *1-69* and *2-70* present in those individuals.

Interestingly, although all the children are homozygous for the GRCh37 (*1-69/2-70*) haplotype without the insertion, three of them have the GRCh38 (*3-64D/5-10-1*) haplotype on at least one chromosome. This implies that there are IGHV haplotypes different from both reference genomes and that are possibly mosaics of the reference genomes. Analysis of these two variants therefore not only confirms their presence in the Platinum Genomes sample but also demonstrates that different configurations are present in the same ethnic population and that many more configurations may exist. This is in line with resequencing efforts that have discovered novel sequences not found in GRCh37 [117, 47, 46, 56], highlighting that the diversity of alternative haplotypes remains largely unexplored.

High copy number variation in the operationally defined *3-30* segment. Among all the segments, *3-30* exhibited the most variation in coverage depth (Figure 2.5). This confirms previous findings that the IMGT alleles found near this region, including those of *3-30/3-30-3/3-33*, often exhibit differences in copy number [121, 88]. The distribution of this variation has not previously been characterized, however. With the application of our pipeline to reads from the operational segment *3-30*, we are able to begin collecting previously unknown quantities, such as the mean and range of copy numbers. This information will also help determine whether copy number is segregating in different subpopulations.

Using pedigree information as a constraint, we reconstructed the diploid configuration of copy number for segment *3-30* in each member of the family (Figure 2.7). Its abundance ranges from zero to four copies on a chromosome. To our knowledge, previous results about the variation in copy number of this segment were not linked to the ethnicity of the individual. Therefore, this may be the first result about copy number variation of *3-30* within a small sample of individuals of European ancestry.

New *7-4-1* allele. With the exception of *7-4-1*, we found exact matches to IMGT alleles amongst all the segments in the Platinum Genomes dataset. In the case of *7-4-1*, none of the assembled alleles exactly matched an existing IMGT *7-4-1* allele. To eliminate the possibility that the allele calls were confounded by reads from pseudogenes, these alleles were determined after applying an extra filtering step to reads that mapped to the *7-4-1* segment

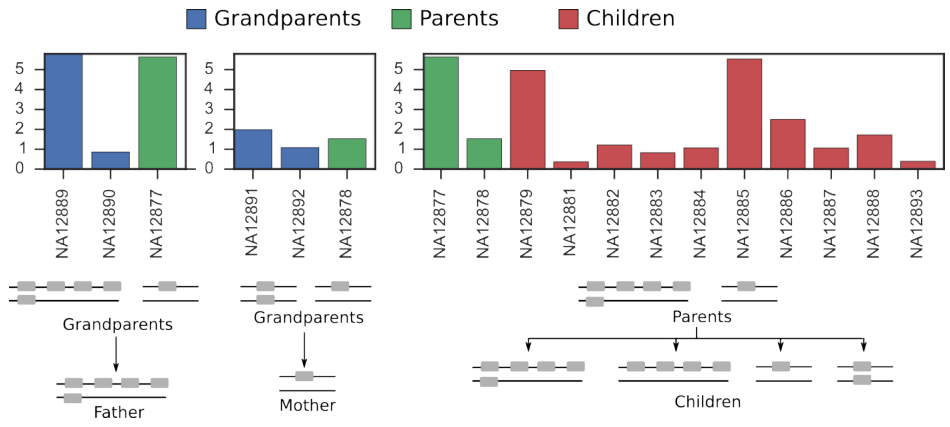


Figure 2.7: Copy number variation in the *IGHV3-30* segment. Below each bar chart of coverage values is a putative reconstruction of the genomic configuration of the individuals in the family. Y axis is normalized coverage.

(Section 2.3). One of these alleles was five nucleotide mutations away from *7-4-1∗04* and present in five individuals (Figure A.7 shows an alignment of the new allele with *7-4-1∗04*). Our finding of an allele that is not in the IMGT database is in line with recent reports of novel alleles found using antibody repertoire sequences [10, 119, 34, 103].

2.3 Methods

The standard naming convention of IGHV genes

IGHV genes are named according to their "family" and genomic location. The families, numbered 1 to 7, comprising genetically similar genes. The segment *6-1*, for example, is in IGHV family 6 and is the first gene in the locus, counting from the centromeric end. Gene names with a suffix "D" denote a duplicate gene, for example *1-69D*, while an appended number, for example 1-69-2, indicates that the gene was discovered subsequent to the original labeling and is located between 1-69 and *2-70*. An allelic variant of an IGHV gene is denoted by a ∗01, ∗02, etc, as in *1-69∗01*, *1-69∗02*.

Hierarchical clustering

Nucleotide sequences for IGHV gene alleles were downloaded from the IMGT database [54]. Only full-length functional alleles were used for clustering. Multiple sequence alignment was performed on each family of alleles using Fast Statistical Alignment with default parameterization [11]. The aligned alleles were then clustered using the `hclust` function in R (method parameter set to "single", although using the "complete" method gives the same result for all families with the exception of family 4, [77]). For all the IGHV families except family

4, operational segments were determined using distance matrices calculated from Hamming distance based on FSA alignment, with gap differences treated in the same way as mutations. Visual inspection of the alignment of family 4 suggested that indels may be important in partitioning the alleles. Hence, a combination of an evolutionary distance "TN93" (based on [113]) and indel distance (number of sites where there is an indel gap in one sequence and not the other) was used to determine the operational segments for family 4.

Genotyping pipeline

Our scripts and example datasets are available at: `https://github.com/songlab-cal/SGDP_IGHV_TRBV`. We assume the WGS data is in bam or sam format [57], with reads already filtered to come from the IGHV locus. For WGS reads aligned to GRCh37, this is chr14: 105,900,000-107,300,000. For reads aligned to GRCh38, this is chr14:105,700,000-106,900,000 (coordinates extend beyond the IGHV locus to be conservative). Bowtie2 [51] is used to map these reads to all functional, full-length IMGT alleles (the same set used for hierarchical clustering). The default Bowtie2 local alignment threshold led to too many multiple matches. Figure A.9 illustrates how we increased this threshold to be more restrictive. Mapped reads are then pooled according to the operational segments described in the Results section. For example, all reads that map to the alleles of *3-30*, *3-30-3*, *3-33* are pooled together. SPAdes *de novo* assembler [7] is run on the pooled reads for each operational segment. The assembled contigs are compared with the IMGT database using stand-alone IgBLAST [124] to determine the closest matching allele, the length of the match, and the number of nucleotide mutations or indels that separate the contig from the closest-matching allele. The read coverage depth of the contig as reported by SPAdes is also recorded for further analysis.

Simulated reads

To test the capabilities and quality of our methods, ART [39] was used to generate simulated Illumina reads from GRCh37 and GRCh38 of lengths 70, 100, and 250 bp, each at coverage depths of 30×, 40×, and 50×. Error profiles of simulated reads and adjustments to default ART parameters are illustrated in Figure A.10 and Figure A.11.

Filtering using mate-pair information

For the Platinum Genomes data, which comprises paired-end reads, we apply an additional filtering step to remove reads from pseudogenes that share a common subsequence with a functional gene. One way to disambiguate a read of a functional gene from one of a pseudogene is to compare the genomic position that its mate maps to. If the mate read maps to a region that is substantially farther from the region the first read maps to (we use a threshold of 1000 bp to be conservative) then there is a chance it comes from a pseudogene and the original read is discarded. Figure A.12 demonstrates that this filtering step eliminates more than half the reads from pseudogenes. Note that as a tradeoff, this

filtering step will in some cases also incorrectly discard reads from duplicates that are located in a different region of the genome. For segments where the starting position relative to the genome is undetermined, no filtering occurs. In the case of the Platinum Genomes data, which is aligned to GRCH37, this means that filtering is not applied to reads from segments *7-4-1*, *5-10-1*, *4-38-2*, *4-30-2*, and *1-69-2*.

Extra filtering step for novel *7-4-1* allele detection

Alleles of *7-4-1* have high nucleotide similarity to subsequences of pseudogenes *7-81*, *7-40*, and *7-34-1*. The mate-pair filtering step above does not apply to *7-4-1* because the Platinum Genomes reads are aligned to GRCh37, which does not contain *7-4-1*. To filter out reads from these pseudogenes for *7-4-1*, we ran stand-alone IgBLAST on reads mapped to segment *7-4-1*. The reads that had the highest match to a pseudogene were removed. The remaining reads were then used as input for SPAdes *de novo* assembler.

2.4 Discussion

With the approach introduced here, we can begin to obtain population-level statistics on the IGHV locus and quantify its variation. Given the small sample size of the Platinum Genomes data, we have focused here on quantifying variation in genes known to vary in copy number. As larger whole-genome sequencing datasets become available, it will be possible to compare IGHV copy number profiles at the population scale. These profiles can then be studied to find correlations between multiple gene segments and to discover new copy number variants. Even with the coarse presence/absence of segment genotypes, we can begin to address basic open questions such as whether there is a minimal number of IGHV gene segments required for a healthy immune system and whether there is a common core set of IGHV genes that are shared by all individuals.

Our study makes clear that read depth information can be used to accurately determine the presence and absence of gene segments. However, complications remain for ascertaining copy number and allelic content to high accuracy. The first complication arises from the cell type on which whole-genome sequencing is commonly performed. The Platinum Genomes data were generated from immortalized B lymphocytes. The IGHV locus in these cell types has undergone VDJ recombination. This rearrangement, which truncates the IGHV locus, confounds the correlation between read coverage depth and copy number of a gene segment. We can see this from the pipeline output, where coverage depth tends to decrease towards the centromeric (*6-1* segment) end of the locus. The extent of this decrease can be quite marked, for example in the case of NA12877, or not noticeable at all, for example in NA12891 (Figure 2.8A; the distribution of read coverage depth of all the individuals is summarized in Figure A.5). If one knew the number of B cell lineages used to prepare the library and the fraction of haplotypes that underwent rearrangement, it is possible to adjust the raw coverage values to reflect actual coverage values (See Appendix A). However, in the case of

the Platinum Genomes data, this information is unavailable. As whole-genome sequencing becomes more widespread, we anticipate that datasets from other cell types will become available and this issue will be resolved.

The second complication is that the majority of whole-genome sequence reads are generated from diploid cells. Because the majority of segments on both chromosomes are of different alleles, the single allele call generated by our pipeline may be composed of sequences from all the alleles present or represent just one of the alleles. Figure 2.8B shows that allele calls can hide the heterozygous state of an individual. Figure A.8 gives further examples of segments that are present as two alleles in the family and for which the allele calls are misleading. This problem could be addressed with an assembler customized to identify allelic variants of short genomic regions (popular assemblers are currently designed for whole-genome assembly). There has been some success in identifying unique alleles using an alternative data type: antibody repertoire sequencing data [10, 48, 34]. However, such studies cannot directly quantify the copy number of an exactly duplicated gene because read abundance in these studies is not correlated with germline gene abundance. Furthermore, the V gene segment is truncated during the genomic rearrangement for producing the antibody coding sequence, so that full-length alleles may not always be obtained from antibody repertoire sequencing data.

We note that there are many existing methods for estimating copy number based on coverage depth using whole-genome sequencing [13, 4, 15, 125, 1]. These methods, however, do not utilize the IMGT database of IGHV alleles or specifically target the IGHV locus, a region with a higher amount of repetitions and duplications than most of the genome. They, therefore, may be prone to biases introduced by targeting the entire genome, which has loci of varying characteristics, rather than targeting a particular region. Additionally, some existing methods [49] intended for whole-exome sequencing may be further biased when introduced to data from whole-genome sequencing.

True determination of IGHV haplotypes must ultimately come from sequencing the 1 Mb region in its entirety and in multiple individuals. Indeed, because the GRCh37 reference is a chimera of three diploid haplotypes [71], there is currently only one true reference haplotype for the IGHV locus. However, the technology to accurately sequence structurally varying regions remains expensive and low-throughput. We can instead take advantage of the increasing availability of whole-genome sequencing datasets and the extensive IMGT database to genotype this locus in a high-throughput manner but at lower genotypic resolution. With this approach, we have found evidence of haplotypes that are mosaics of reference genome configurations or that are transitional between them. The existence of these haplotypes further indicates that our approach of representing the locus in terms of the copy number of a reference set of segments is better suited to cataloging variation in this locus than full sequences of the IGHV locus with annotated breakpoints.

The fundamental strategy applied here is not specific to the IGHV locus. Reads from whole-genome sequencing datasets can similarly be used to characterize other gene families and in other species, where the genes are of comparable length and a similar level of diversity. Some examples include T cell receptor genes and olfactory receptor genes. The analysis of

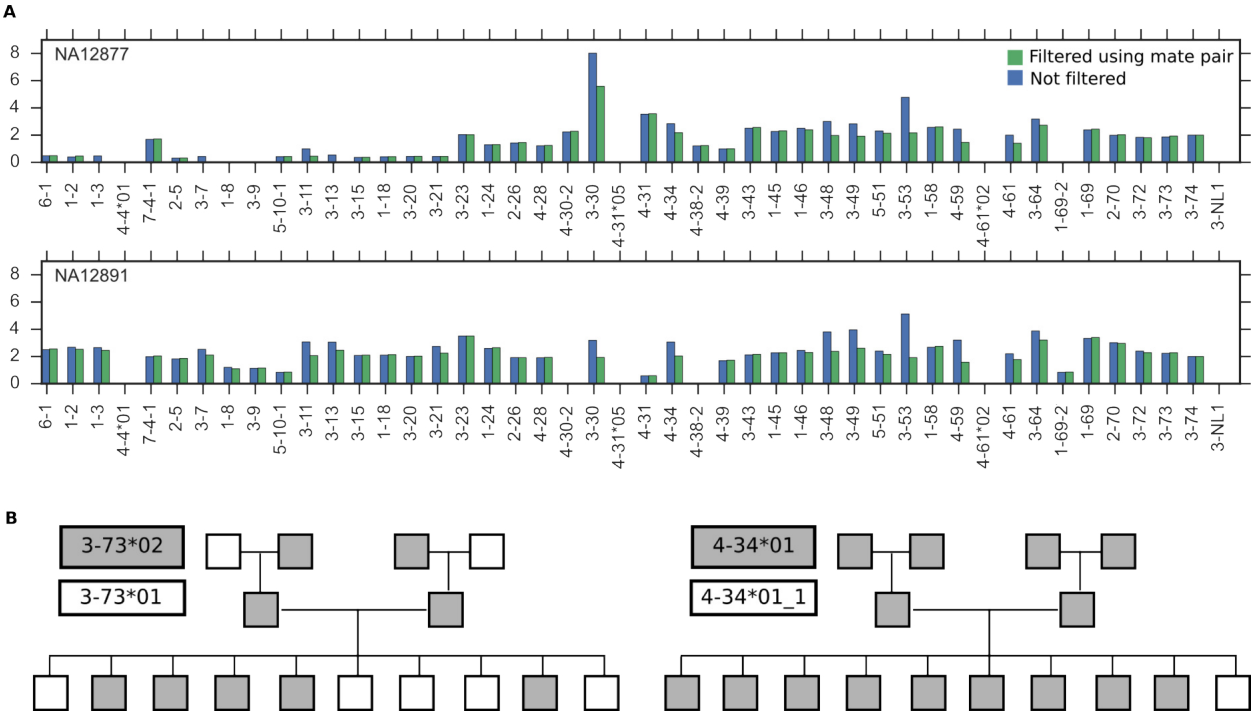


Figure 2.8: Complications arising from cell type and diploidy in Platinum Genomes dataset. (A) Individuals differ in the uniformity of coverage over segment types when DNA is sequenced from B lymphocytes. Y axis is normalized coverage. (B) Allele calls color-coded by allele and arranged by family tree. For *3-73*, at least one of the parents must be heterozygous. For *4-34*, the singleton allele in one child indicates that one parent and its parent is heterozygous or that the allele call is incorrect (*4-34*01_1* is a variant that is not in the IMGT database and is one nucleotide mutation away from the *4-34*01* allele).

whole-genome sequencing data thus need not be restricted to single nucleotide variants, but can also be used to study regions exhibiting copy number variation.

Chapter 3

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

This chapter is joint work with Shishi Luo, Heng Li, and Yun S. Song which appears in *Life Science Alliance* [65].

3.1 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 ([111, 117]). Yet copy number variation remains challenging to quantify and analyze. Nowhere is this truer 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 [79]. Two gene families that are of particular biomedical relevance but for which variation is not well characterized are the immunoglobulin heavy variable (IGHV) family, a 1 Mb locus located on chromosome 14 [71, 121], and the T cell receptor beta variable (TRBV) family, a 500 kb locus located on chromosome 7 [96]. Both regions 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 [76]. 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 [33, 24, 14]. Indeed, the genes comprising the IGHV and TRBV loci are distant paralogs and are believed to derive from a common ancestral locus in a vertebrate contemporaneous with or predating jawed fishes [33, 24, 14]. That these two loci share genomic features and evolutionary origins make them an ideal system for a comparative study in gene family evolution.

Here we present the largest investigation to date of genetic variation in the IGHV and TRBV loci using short-read whole-genome sequencing data. We apply a customized genotyping pipeline (based on [64]) to data from the Simons Genome Diversity Project (SGDP) [67], which performed whole-genome sequencing of a globally diverse sample of human individuals from over a hundred populations. Such characterization of population-level genetic variation in the immune receptor loci sheds light on how the two loci evolved from their common origins. Quantification of variation is also needed in the burgeoning field of computational immunology [35, 95], where the relative abundances of germline variants will help in other applications such as genome-wide association studies, measuring linkage disequilibrium, and determining clonal lineages from VDJ sequences. For example, previous work demonstrates that the V genes may contribute a significant proportion of the CDR3, and oftentimes lineages with conserved D and J genes must be distinguished using V gene information [62]. Past methods for CDR3 determination have included integrating over all possible V genes when information was lacking, and taking population-wide frequencies into account would likely improve the accuracy of such methods [78]. Additionally, the common copy number polymorphisms we find in our data agree with what has previously been documented, and the most frequent allele we report for each gene segment corresponds to the first or second allele ($*01$ and $*02$, respectively) recorded for that gene segment in the IMGT database [37]. We emphasize, however, that the results from this genotyping are used purely for aggregate measures of sample-level variation. Our method is not intended to be used to accurately genotype individual genomes.

3.2 Results

A brief note about gene nomenclature: for the bioinformatic analysis, it was necessary to group together gene segments that are operationally indistinguishable but which have distinct names because they occupy physically different positions in the genome. Our departure from this standard nomenclature is detailed in the Methods section 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 (SNV) or a gene copy number variant (CNV). In line with the usage in the immunogenetics community, 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*. We use *haplotype* to refer to the set of operationally distinguishable gene segments that are inherited from a single parent.

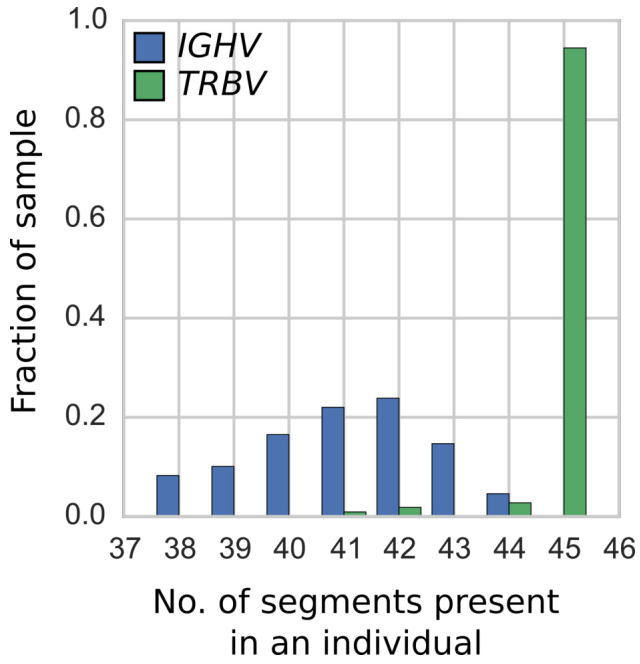


Figure 3.1: Histogram of the number of gene segments in an individual. The results are based on the IGHV (blue) and TRBV (green) segments present in each of the 109 individuals from blood and saliva samples. The number of operationally distinguishable IGHV gene segments shows greater variation than the number of TRBV gene segments. Figure B.9 shows a histogram of the number of TRBV gene segments in the full set of 286 individuals.

Copy number variation

In general, gene duplication/deletion appears to have occurred more frequently 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 3.1, Figure B.9). Using our per-segment copy number estimates and hierarchical clustering (see Within-species analysis, Supplementary Information Figures 1 and 2 of [65]), we identified locus-wide copy number haplotypes, some of which have been previously reported (Figure 3.2 and 3.3). To be conservative, we restricted our figure results to polymorphisms that either involve at least two operationally distinguishable gene segments or involve a single gene segment with high levels of copy number variation. Several IGHV genes (*IGHV7-4-1*, *IGHV4-4*, *IGHV4-30-4*, *IGHV4-59*, and *IGHV4-61*) had unusual read-coverage profiles and, to be conservative, were not included in the CNV calls. In contrast, TRBV genes had predominantly well-behaved read coverages and were two-copy per individual, resulting in a more complete list of CNVs in TRBV (Figure 3.3).

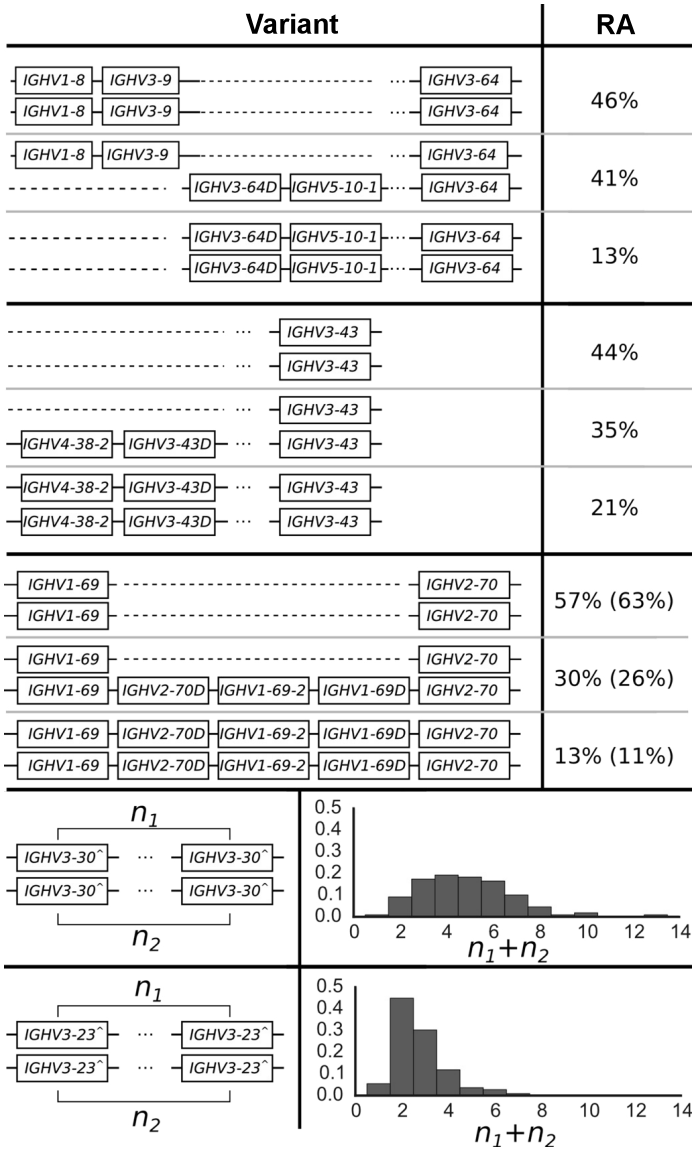


Figure 3.2: The distribution of IGHV copy number polymorphisms reliably called in our sample. Schematics in the left column show the polymorphisms while the right column displays the relative abundance (RA) in the sample of 109 individuals. For the polymorphism involving *IGHV1-69*, we also show the relative abundances in the full sample of 286 individuals in parentheses. This is because *IGHV1-69* and *IGHV2-70* are located in the J-distal part of the IGHV locus, making them less likely to be affected by VDJ recombination. Unlike the IGHV polymorphisms that are closer to the J region, we saw negligible differences in copy number estimates for these gene segments in the saliva versus cell-line samples. 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*.

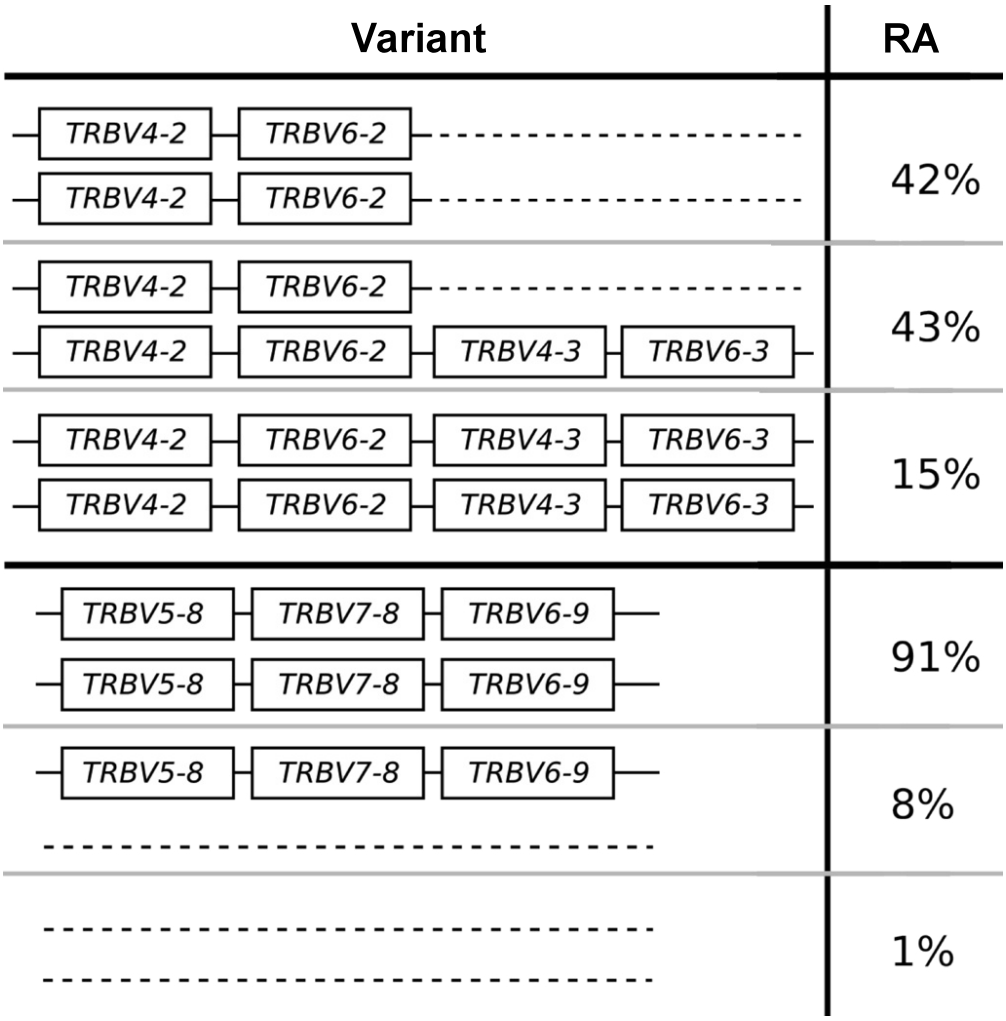


Figure 3.3: The distribution of TRBV copy number polymorphisms reliably called in our sample. Schematics in the left column show the polymorphisms, while the right column displays the relative abundance (RA) in the full sample of 286 individuals. Our data informs the copy number of these genes, while the genomic configuration is our best estimate based on previous studies. The insertion of *TRBV4-2*, *TRBV4-3* and *TRBV6-2*, *TRBV6-3* is a frequent polymorphism also found in previous studies [110, 105, 129]. The polymorphism involving *TRBV5-8*, *TRBV7-8*, and *TRBV6-9* was identified by first clustering using *TRBV5-8* copy number estimates alone, and then noticing that such a clustering also induced a clear-cut partition of the copy number estimates for *TRBV7-8* and *TRBV6-9*. See Appendix.

Lack of geographical associations. We considered grouping individuals according to the geographic regions defined by SDGP, namely, Africans, West Eurasians, Central Asians-Siberians, East Asians, South Asians, Oceanians, and Native Americans. In the majority of cases, we found that the distribution of copy number variants within a geographic region is consistent with the global distribution (Figures B.11 and B.12). The two exceptions are: (i) the polymorphism involving *IGHV1-69*, where the duplication/insertion variant is the major variant among genomes sampled 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 genomes sampled from the Americas, 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. We note, however, that the sample sizes for the IGHV analysis of East Asians, Oceanians, and Native Americans do not have suitable statistical power, and are included for comprehensiveness and illustrative purposes.

No correlation between copy number polymorphisms. We found effectively no correlation between copy number polymorphisms in either IGHV or TRBV (Figures B.13 and B.14). The average value of R^2, the square of the Pearson correlation coefficient, between segments in the different polymorphisms is 0.021 for the IGHV gene segments (Figure 3.2) and 0.004 for the TRBV gene segments (Figure 3.3). Thus, the polymorphisms are essentially independent, and we can estimate the number of copy number haplotypes in the two loci. From Figure 3.2, 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 ($2 \times 2 \times 2 \times 7 \times 4$), assuming independence between the common copy number polymorphisms. The analogous calculation from Figure 3.3 for TRBV leads to only a handful of haplotypes (2×2). We note that this result is not meant to be taken literally. Rather, the orders of magnitude difference between our estimates for IGHV haplotypes compared to TRBV haplotypes strongly suggests that the two loci have undergone different rates of gene duplication and deletion.

SNV and allelic variation in two-copy gene segments

Having quantified copy number variation of gene segments across the two loci, we sought to compare nucleotide variation while minimizing the confounding factor of copy number variation. A gene segment with a 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. For this reason, we compared single-nucleotide and allelic variation in IGHV and TRBV gene segments that have two copies in the vast majority of individuals in our sample and for which there is minimal read-mapping ambiguity (11 such IGHV segments, 40 TRBV segments, see Supplementary Information Figures 1 and 2 of [65]). We will refer to such gene segments as "two-copy" for short. In this context, single nucleotide polymorphisms (SNPs) are meant to refer to nucleotide positions that are polymorphic when compared across

	IGHV	TRBV
Average bp difference per pairs of alleles	4.1%	5.0%
Average number of SNPs per segment	1.7%	1.9%
Fraction of novel alleles out of all observed alleles	12/28(48%)	40/99(40%)

Table 3.1: 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. To calculate the average base pair difference per pairs of alleles, for each segment we computed the average base pair difference between all pairs of alleles, and then averaged over all segments.

individuals in our sample, while a single nucleotide variant (SNV) is a specific genetic type occurring at a SNP. This is in contrast to a "novel allele", which refers to a sequence of nucleotides that do not exactly match any known allele in the IMGT database. In addition to restricting our single-nucleotide and allelic analysis to two-copy gene segments, we were also conservative in how we called these variants: an allele or SNV is called only if it is present in two or more individuals. To be clear, the only analysis that is limited to 11 IGHV genes (as opposed to 40 TRBV genes) is the allelic/SNP variation analysis.

IGHV and TRBV have comparable levels of nucleotide diversity in two-copy genes. 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 3.1, Figure 3.5). If anything, the TRBV two-copy gene segments exhibit greater single-nucleotide and allelic diversity, given the higher number of SNVs and average base pair differences between the 109 individuals. We find that on average, IGHV two-copy gene segments have 1.7 SNVs, as opposed to 1.9 SNVs per TRBV two-copy gene, which is similar to previous work reporting roughly 2 SNVs per gene [66, 110]. Our slight underestimate of this value is reasonable given that we restrict our analysis to two-copy genes. That TRBV exhibits greater or comparable diversity is seemingly surprising, because if allelic diversity is estimated by taking the average number of alleles per gene segment as per the IMGT database, without regard to the segment's copy number, operationally distinguishable IGHV gene segments have an average of 5 alleles while TRBV gene segments have an average of 2 alleles. This discrepancy in the two ways of estimating sequence variation does not seem to be due to an under-representation of TRBV alleles in the IMGT database relative to IGHV alleles: the fraction of putative novel alleles called in our sample is similar between the IGHV and TRBV gene segments (Table 3.1, third row). However, if high copy-number segments were included in this allelic diversity analysis, then the per-segment allelic diversity for the IGHV locus would likely be higher than the observed diversity for two-copy segments, as suggested by the results of [103]. This discrepancy could indicate that our observation of elevated nucleotide diversity in two-copy gene segments may not hold for the

IGHV and TRBV loci as a whole; it could be that restricting to two-copy gene segments filters out IGHV genes with higher levels of nucleotide diversity, which could have resulted from relaxed selective pressure in higher copy-number gene segments. 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 Tables B.1 and B.2.

Putatively novel alleles. We called 28 IGHV alleles, of which 12 are putatively novel and 97 TRBV alleles, of which 38 are putatively novel. Of these novel alleles, it is notable that 5 IGHV alleles and 12 TRBV alleles appeared at least 10 times in our sample (we count homozygous alleles as appearing twice, Table B.4). Some of these novel alleles like *IGHV1-45∗02_ga123GR*, *TRBV10-1∗02_gt234E_*, and *TRBV12-5∗01_cg27HD* (see Section B.7) are present in high frequency across all geographic regions. That these novel variants are comprehensively present supports existing evidence that the databases of IGHV and TRBV alleles are not yet complete [103, 34, 22, 36].

SNV and allelic variants private to geographic regions. A SNV that is private to a geographic region indicates that individual(s) all from one region have a base pair that differs from the base pair of all other individuals at that site. Alternatively, an allele that is private to a geographic region indicates that an entire allelic sequence is specific to individual(s) from that region. We found 5 SNVs in the 11 two-copy IGHV gene segments that are private to a single geographic region, and 14 such variants in the 40 two-copy TRBV gene segments (Table B.5). These variants are not rare: a majority of them are present at greater than 10% frequency in the geographic region to which they are exclusive, with the extremes being as high as 42%. For both loci, the geographic region of Africa had a disproportionate share of such variants: of the 5 IGHV SNVs that were private to a geographic region, all 5 were private to Africa, and of the 14 SNVs exclusive to a region for TRBV, 10 (71.4%) were private to Africa (Table B.5). This particular feature of samples from the Africa region is also apparent in our allelic variation analysis. Of the 28 IGHV alleles we called, 4 out of 4 private alleles were private to Africa. Similarly, of the 97 TRBV alleles we called, 10 out of 14 (71.4%) private alleles were private to Africa. These findings of higher levels of diversity primarily in Africa are consistent with prior studies [66, 20, 42, 59, 91, 114, 115, 116, 129] and with the percentage of exon-located single-nucleotide variants that are private to Africa across the entire genome (72.1%). For a complete table of SNVs and alleles private to a particular region, see Table B.5.

Geographical clustering of TRBV haplotypes. To investigate whether genetic variation at immune receptor loci exhibits geographical structure, we applied multidimensional scaling to the reconstructed phased TRBV haplotypes from 286 individuals for each two-copy gene. Figure 3.4 illustrates our result, where each point corresponds to an individual. Figures B.15, B.16, B.18, and B.17 show results from applying multidimensional scaling to individuals just from pairs of geographic regions.

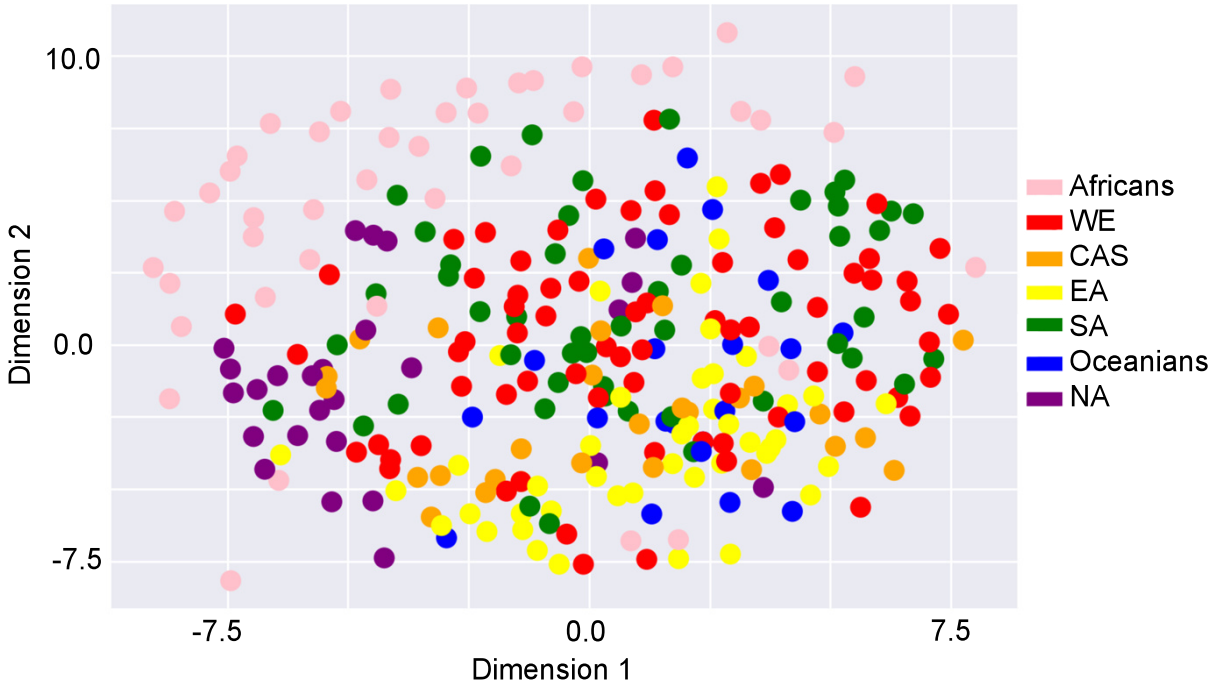


Figure 3.4: Multidimensional scaling of TRBV alleles. Based on 286 individuals from all populations (including all DNA source types) in the SGDP dataset. Each point corresponds to an individual and is colored by the corresponding geographic region defined by SDGP: Africans, West Eurasians (WE), Central Asians-Siberians (CAS), East Asians (EA), South Asians (SA), Oceanians, and Native Americans (NA). Multidimensional scaling was performed in Python using the `manifold.MDS([n_components, metric, n_init, ...])` function from the sklearn.manifold module. The data fit by the model uses the Euclidean distance between x_i and x_j where the mth entry in vector x_i is the copy number of allele m in individual i, taking possible values 0, 1, or 2.

As shown in Figure 3.4, we observed the clearest separation between the African population and the rest of the populations, a trend that is also apparent in the pairwise plots (Figure B.15) and is related to our aforementioned finding that African individuals tend to have the most alleles private to one region. There are a few African individuals who are exceptions to this pattern. Specifically, Masai-1 from Kenya and Saharawi-2 from Morocco consistently cluster more closely with Eurasians.

While distinction amongst the other populations is not immediately obvious from Figure 3.4, every pairwise comparison with Native Americans showed reasonably clear separation from the other populations (Figure B.16), which may be due to the reduced genetic diversity of Native Americans compared to that of other populations [118]. Additionally, the individuals from Central Asia-Siberia and South Asia were fairly separable, although

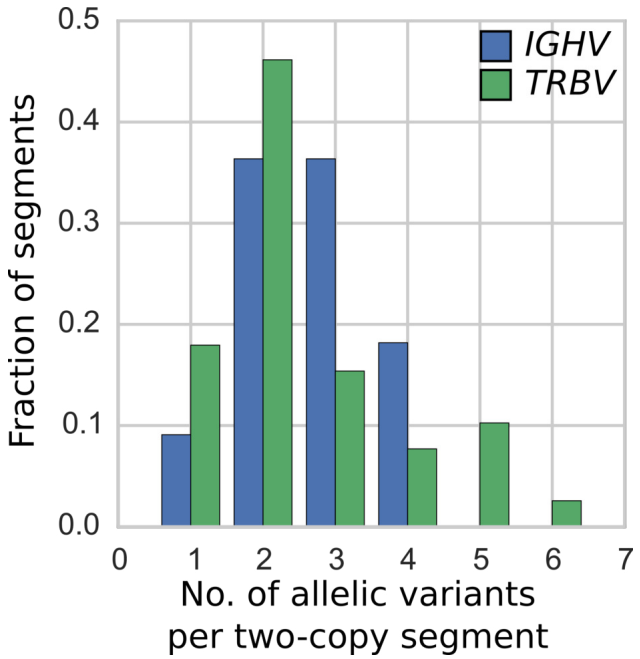


Figure 3.5: The number of alleles in the 11 two-copy IGHV (blue) and 40 two-copy TRBV (green) segments. We report an allele only if it is found in at least two out of the 109 genomes from blood and saliva samples. The two distributions are not statistically significantly different (p-value of two-sample Kolmogorov-Smirnov test between the blue and green distribution is 0.97).

the degree of distinction is less prominent compared with those discussed above. Comparisons demonstrating significant overlap include Central Asia-Siberia versus East Asia, West Eurasia versus Central Asia-Siberia, and West Eurasia versus South Asia, which is expected given previous reports of high gene flow between Europe and Asia [90]. Given these results, we would expect high fixation index values between each subpopulation and Africa/Native America, and lower fixation index values otherwise, which is indeed what we find (Table B.3).

General variation patterns suggest distinct evolutionary dynamics

Our analysis of all functional operationally distinguishable gene segments in the two loci indicates more gene duplication/deletions in IGHV than in TRBV (Figure 3.1). In contrast, the observed level of nucleotide diversity within gene segments—as measured by the amount of sequence variation per gene segment in two-copy genes—seems to be slightly higher in the TRBV locus than in the IGHV locus (Table 3.1, Figure 3.5). If the rate of sequence diversification were indeed higher in TRBV than in IGHV, we would expect the IGHV gene family to comprise genes that are more similar to each other on average than the TRBV gene family. This holds true for the genes found in the IMGT annotated gene table for humans.

For all pairs of functional genes, we measured between-segment diversity as the pairwise global alignment score (see Within-species analysis for details) between gene segments, which gives significantly higher scores for the IGHV genes, indicating more mismatches and gaps between TRBV genes. Using this same set of human genes and an annotated dog reference genome (also curated on IMGT), we performed a similar analysis in IGHV and TRBV gene families between humans and dogs and found similar results.

Given the larger diversity among TRBV genes between these two species, we then looked at amino acid diversity in IGHV and TRBV gene families within each of thirteen vertebrate species (curated at vgenerepertoire.org), including five primates, six non-primate mammals, one reptile, and one fish (Figure 3.6, Appendix B); we again found a similar pattern for each species. The amino acid diversity for each species was calculated between the IGHV genes and between the TRBV genes in that species' reference genome. For all these species, we found that the IGHV gene segments have substantially lower within-species diversity (about 44%) than the reference TRBV gene segments (about 60% within-species diversity; Figure 3.6). We also observed less homology between species for the IGHV gene family compared to the TRBV family (Figures B.10), which together with the aforementioned lower diversity in IGHV, suggests that IGHV homologs that are shared between species are deleted more frequently than TRBV homologs. This is consistent with our finding that gene duplication and deletion occur more frequently in the IGHV locus. It is possible, however, that rather than being erased, some genes accumulate sufficient amounts of nucleotide changes that cause them to appear as an entirely new gene.

3.3 Methods

Gene nomenclature

The following sets of gene segments were considered operationally indistinguishable (often more than 95% nucleotide similarity) 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

Whole-genome shotgun sequencing reads were collected in a previous study, the Simons Genome Diversity Project [67]. Briefly, 300 genomes from 142 subpopulations were sequenced to a median coverage of 42x, with 100 base pair paired-end sequencing on the Illumina HiSeq2000 sequencers. The reads from 286 of these genomes were mapped to the set of functional alleles (IGHV or TRBV), where our definition of functional is according to the IMGT database annotations [37]. 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 [67]. For the TRBV locus, we used the full set of 286 genomes, unless otherwise stated. Note: we only had access to 286 genomes of the 300 genomes: 300 minus the 14 individuals with labels SS60044XX.

Data availability

The raw data for 279 genomes are available through the EBI European Nucleotide Archive under accession numbers PRJEB9586 and ERP010710. For additional 21 genomes (designated by code Y in the seventh column of Supplementary Data Table 1 in [67]), data are deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001001959. The set of filtered mapped reads can be found at `https://github.com/songlab-cal/SGDP_IGHV_TRBV`.

Read mapping/filtering

For the results above we used reads mapped to a list of functional IGHV and TRBV (from the online IMGT database [37]). The disadvantage of this procedure is that reads from highly similar pseudogenes and orphon genes may get mixed with reads from functional genes (Figure B.20). Thus, 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. We refer the reader to the full details of the filtering steps in the Section B.2.

Copy number calls/contig assembly

After read filtering, we have, for each individual, a set of reads binned by operationally distinguishable segment. We next run the assembler Spades [7] to construct a contig for each segment to obtain:

1. kmer coverage for the segment in that individual
2. A first estimate of the nucleotide sequence of the individual's gene

For example, for a fixed individual, the script we execute to assemble the contig for *IGHV6-1* is:

spades.py –k 21 –careful –s IGHV6-1.fastq –o contigs/IGHV6-1

The choice of kmer of size 21 is because it was the longest kmer that ensured successful contig construction for our 100 bp reads at around 40 coverage depth. The kmer coverage is then converted to per-base coverage, scaled to account for the trapezoidal shape of the read

coverage profile, and then normalized by the individual's genome-wide coverage to obtain a point estimate for copy number (details of calculation in Section B.3).

Haplotype phasing and allele/SNV calls

The contigs and reads for two-copy segments were analyzed for allelic and SNVs by phasing these segments for each individual. Because the assembly step in the pipeline produces only one contig, we reconstructed the two distinct allelic sequences on each chromosome through additional steps, which are as follows:

1. Mapped the filtered set of reads to the contig constructed via the customized pipeline using bowtie2 *–local –score-min G,20, 30*.

2. The results from Bowtie2 were fed to GATK [72] for variant calling, producing VCF files identifying polymorphic sites, using the HaplotypeCaller with parameters *-ploidy 2 -stand_call_conf 30 -stand_emit_conf 10*.

3. The variants from GATK were then phased using HapCUT2 [30]. Procedures for handling instances when HapCUT2 failed are explained in Section B.6. To be conservative, we kept only the alleles found in at least two different individuals.

3.4 Discussion

The analysis of gene families remains a technically challenging task in modern genetics. Here, we have made major inroads in quantifying sample-level variation in gene segments in the IGHV and TRBV gene families. We have uncovered patterns of variation that hint at the evolution of the two gene families as well as allelic variants that may be associated with diseases specific to a geographic region. Our analysis suggests that the IGHV gene family has experienced more frequent gene duplication/deletion relative to the TRBV gene family over macro-evolutionary time scales. The lack of geographical associations for the majority of common copy number polymorphisms in our sample suggests 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. However, we found a number of alleles in both gene families to be private to a particular region and at non-trivial frequencies. Such allelic variants may be promising candidates for investigating genetic variants that are beneficial against infectious diseases endemic in a geographic region. These differences in IGHV and TRBV may be associated with the different functions of B cells and T cells, particularly the latter's interaction with major histocompatibility complex molecules, which itself is complex and highly variable.

Our analysis of copy number variation has practical implications for germline IGHV haplotyping: approaches for cataloging variation by sequencing the 1 Mb locus in full [109, 121] 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 haplotypes, 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 (Table B.6). These results provide quantitative support for the conjecture made by Li et al. [58] 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 (Appendix B). Furthermore, immunoglobulins undergo genetic modification via somatic mutation. The analysis of the antibody repertoire may, therefore, need to be customized to each individual, as suggested by others [22].

Many challenges remain in genotyping complex and variable regions such as IGHV and TRBV. Our approach of using short-read data has a major advantage in being scalable to large sample sizes, allowing population frequencies to be calculated. However, other approaches may be more appropriate if the goal is to genotype a single individual at base-pair resolution, rather than a large set of individuals at a coarser resolution. Another challenge is measuring the rate of nucleotide substitution in IGHV genes, which requires distinguishing between mutations on paralogous regions from true allelic variation. We have adopted a conservative approach here, restricting our calculation to 11 IGHV genes which we are confident are two-copy. However, these 11 genes may not be representative of all the regions of IGHV that are not subject to copy number variation. An approach that can identify larger tracts of IGHV that are structurally conserved across hundreds of individuals will give a better estimate of the nucleotide substitution rate.

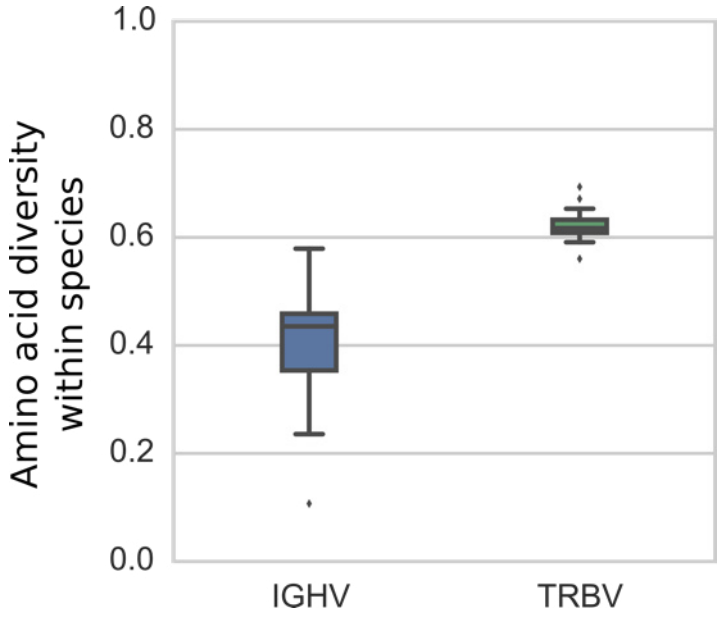


Figure 3.6: Box plots for pairwise diversity between IGHV segments and TRBV segments within a species, averaged across thirteen vertebrate species. For each species, pairwise alignments of all pairs of IGHV segments and all pairs of TRBV segments were performed using ssw [128], an implementation of the Smith-Waterman algorithm [108].

Chapter 4

A fast machine-learning-guided pipeline for SWGA

This chapter is joint work with Matthew Mitchell, Yun S. Song, and Dustin Brisson.

4.1 Introduction

The rapidly expanding field of population genomics is transforming our understanding of how evolutionary forces shape genomic diversity among a wide range of species [81]. In microbial systems, in particular, population genomic studies are increasingly feasible due to the minimal cost of sequencing small genomes [53, 106, 92, 85]. These studies can provide us with the keys to understanding the origins of adaptive traits, mapping expansion, and migration patterns, and understanding epidemiology through the lens of evolution. A principal obstacle to sequencing specific microbial genomes from natural samples is isolating the target microbial DNA from the DNA of contaminating organisms [70]. Although laboratory culture is the standard practice, the overwhelming majority of microbes cannot be cultured and direct sequencing is problematic as the microbial genome constitutes only a minuscule fraction of the total DNA [104, 50, 31]. Thus, a primary hindrance to collecting populations of microbial genomes is the lack of an innovative, cost-effective, and practical method to collect sufficient amounts of target microbial genomic DNA with limited contaminating DNA.

Several technologies have been developed and utilized to overcome this obstacle including genome capture, single-cell sequencing, and selective whole genome amplification (SWGA) [68, 8, 55]. Of these, SWGA is the most inexpensive, flexible, and shareable culture-free technology [97]. SWGA, which takes advantage of the inherent differences in the frequencies of sequence motifs (k-mers) among species such that primer sets bind often in the target genome but rarely in the contaminating genomes, can be used to selectively amplify the target microbial genomes using Φ29 multi-displacement amplification technology [25, 55]. The Φ29 DNA polymerase is strand displacing and amplifies DNA from primers with high

processivity (up to 70-kbp fragments) and is 100 times less error-prone than Taq, making it ideal for genome amplification prior to sequencing [25, 86, 6]. By coupling Φ29 amplification with selective priming, researchers can selectively amplify a target microbial genome, thus separating the metaphorical baby (target microbial genomes) from the bathwater (off-target DNA from vectors, hosts, or other microbes). SWGA has proven to be a powerful and cost-effective tool for researchers looking to generate genomic data for microbial systems. Effective SWGA protocols have resulted in next-generation sequencing (NGS)-ready samples that are enriched for specific target microbial genomes and have been used to address biologically important questions in several microorganisms, including *Mycobacterium tuberculosis* [18], *Wolbachia spp* [55, 18], *Plasmodium spp* [112, 38, 84, 23, 63], and *Wuchereria bancrofti* [107].

The most recent SWGA development pipeline (`swga`) improved on the concept and existing tools available for SWGA primer selection [18]. Whereas the first SWGA tool only used differential binding ratios of k-mers and melting temperature to build primer sets [55], `swga` incorporated a larger *a priori* set of optimality criteria to use when selecting both individual primers (primer binding frequency, improved melting temperature, evenness) and potential primer sets (evenness, primer binding site density on the target genome) [18]. Specifically, data from this study revealed that primer sets generated with `swga` that optimized primer binding site density on the target genome, along with binding site evenness as a secondary factor, yielded the most consistent and useful results. While this program significantly improved upon available SWGA development tools at the time, it suffers from several drawbacks. First, data generated by `swga` uses only marginally-effective optimality criteria to evaluate individual primers and primer sets due to the very limited understanding of the characteristics that result in effective amplification. While primer binding site density and evenness seem to be broadly important, there is considerable variation in amplification success among the primer sets chosen using these criteria. Thus, there are likely novel primer characteristics correlated with efficient selective amplification that are not currently being considered during the process of primer selection. Second, `swga` uses a computationally-expensive algorithm to search for sets of primers that maximize the optimality criteria described above, evaluating no more than 1-5 million viable sets, which are not selected in any optimal way. This is usually only a very limited proportion of all potential primer sets, but can still take time frames that are unreasonable for research projects due to the computational inefficiency. This process of evaluation could be vastly improved by a more well-informed objective function and by pruning unpromising search paths.

The issues just described can make it difficult to develop an effective protocol to amplify the genome of a specific microbial species, and as such, SWGA protocol development is costly in both time and resources. We focused on three ways in which to improve upon current state-of-the-art methods: (*i*) to incorporate active learning and machine learning for modeling primer and primer set efficacy, (*ii*) to incorporate novel features, including thermodynamically-principled binding affinities, (*iii*) to increase computational efficiency, particularly by multiprocessing as much as possible and caching computationally expensive information. This chapter will first present SOAPswga (Selective Optimized Amplifying Primers for selective whole genome amplification), a fast SWGA optimization software that

focuses on these delineated goals. After discussing the pipeline in-depth, we demonstrate its application to *Mycobacterium tuberclosis* from human blood, benchmarking results with previous results with those from [18]. In addition to our SWGA results, we also discuss the methods and results from a set of experiments on plasmid DNA, in which we perform rolling circle amplification (RCA) using individual primers. These results provide insights into the efficacy of individual primers and form the basis on which we filter out low-amplification primers in our pipeline.

4.2 Methods

Here we first discuss each step of the SOAPswga pipeline in-depth, followed by details of how the evaluation function for candidate primer sets was obtained. In the latter portion of this section, we describe experiments conducted in order to understand the fundamental properties of binding between one single primer and one genome. This simplest case allowed us to develop a model for predicting individual primer effectiveness, which is pivotal to the filtration step of the SWGA pipeline and the design of primer sets targeting *Mycobacterium tuberculosis*.

SOAPswga pipeline

The overall pipeline for proposing primer sets takes in a set of genomes which we would like to amplify (target genome(s)) and would not like to amplify (off-target genome(s)), where the goal is to find sets of primers which will effectively and evenly amplify the target genome(s) and not the off-target genome(s). The process is broken into four main stages: 1) preprocessing of locations in the target and off-target genome of all motifs in the target genome, 2) filtering all motifs in the target genome based on individual primer properties and frequencies in the genomes, 3) scoring the remaining primers for amplification efficacy using a machine learning model, and 4) searching and evaluating aggregations of primers as candidate primer sets. See Figure 4.1 for a diagram of the pipeline. Below we discuss each step more carefully.

Step 1. k-mer preprocessing: identifying all 6-12mers in the target genome. The primary step in the program identifies the k-mers of length 6 to 12 in the target genome, which serve as possible candidate primers for downstream steps. The counts of these k-mers in the target and off-target genome(s) are computed using `jellyfish` ([69]), a fast, parallel k-mer counter for DNA. An h5py file is then created for storing primers and their respective locations for each genome (or chromosome). This entire preprocessing step is parallelized, and we have provided pre-computed files for *Mycobacterium tuberculosis* as well as human. If parameters of the pipeline are changed, this step will not need to be re-run.

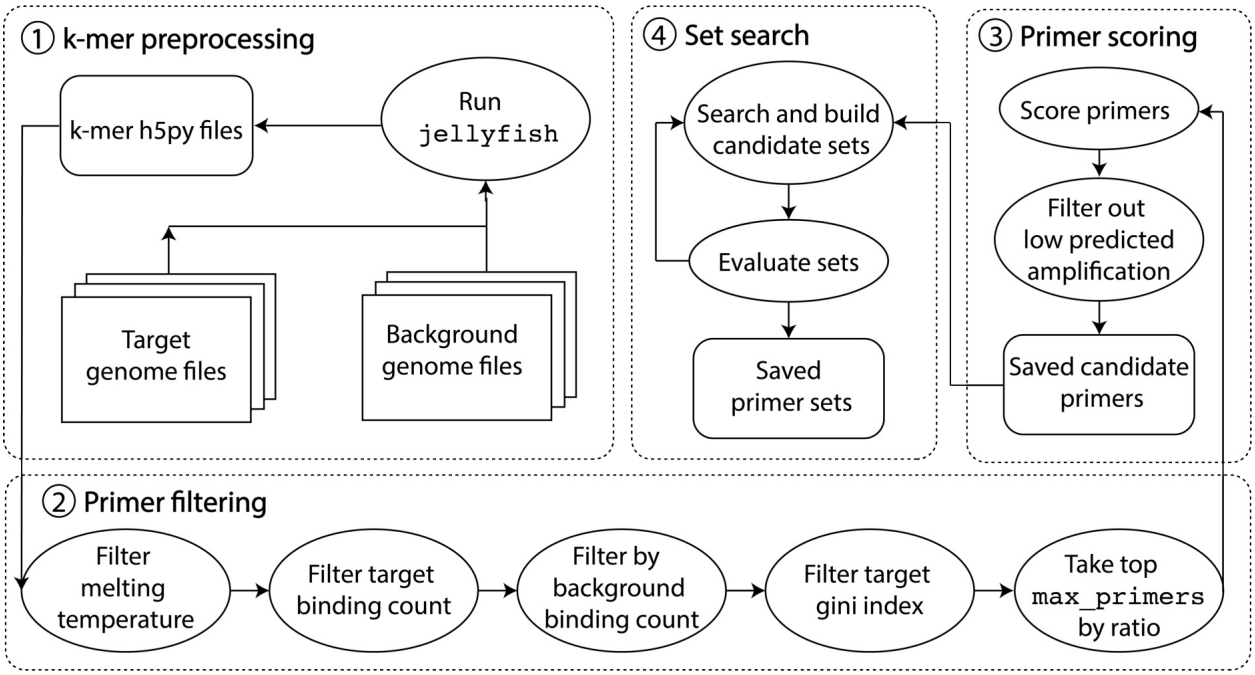


Figure 4.1: Overview of the SOAPswga pipeline. The process is broken into four main stages: 1) preprocessing of locations in the target and off-target genome of all motifs in the target genome, 2) filtering all motifs in the target genome based on individual primer properties and frequencies in the genomes, 3) scoring the remaining primers for amplification efficacy using a machine learning model, and 4) searching and evaluating aggregations of primers as candidate primer sets.

Step 2. Candidate primer filtering: excluding candidate primers which may misprime. In this step, we filter out candidate primers from the set of all motifs in the target genome based on having certain properties of each primer (as suggested by [89]). These are as follows:

1. Melting temperature: the melting temperature must be within the standard temperature range of `min_tm` (default 15°C) and `max_tm` (default 45°C) as established in [55]. These temperatures are computed based on [5].

2. Self-dimer: candidate primers that could possibly form self-dimers are eliminated at this stage. This is estimated by the longest common subsequence between the candidate primer and its reverse complement being greater than `default_max_self_dimer_bp` (default 4).

3. Number of consecutive runs of a single base pair: primers with runs of 5 of five or longer are eliminated.

4. GC content: The GC content for each primer is maintained to be within `min_GC` and `max_GC` (default 0.375 and 0.625 respectively). Three or more G or C's are avoided in the last five and three base pairs of the 3′-end of the primer.

5. Di-nucleotide repeats: di-nucleotide repeats of 5 or more are avoided.

6. Binding frequency: binding frequency is computed as the number of exact matches of a primer in the genome, normalized by the total genome length. Primers that bind too sparsely to the target genome (lower than parameter `min_fg_freq`) or too frequently to the off-target (higher than `max_bg_freq`) are removed.

7. Binding evenness: the evenness of binding is calculated by finding the Gini index of the distances between each primer binding site on the target, and primers with Gini indices higher than `max_gini` are removed.

While the melting temperature filter is more a requirement of the experimental procedure, most of these filters are done in an effort to reduce mis-priming (binding to an undesirable location, e.g. the off-target genome or another primer). The computing of the frequencies of these candidate primers in the genomes is critical for estimating the number of possible binding positions in the target and off-target genomes, the former of which we would like to maximize and the latter of which we would like to minimize. Binding evenness as measured by the Gini index of the gap distances is important for downstream analysis such as short-read assembly and copy number estimation.

Finally, primers are ranked by the ratio of the binding frequency in the target genome(s) to the binding frequency in the off-target genome(s) and those primers with the highest ratio are identified for downstream use (by default, this currently identifies the top 500 primers, and is modifiable via the `max_primers` parameter).

Step 3. Amplification efficacy scoring: predicting individual potential strength to amplify. In this step, before we optimize over the combinatorial space of primer sets, we score each individual candidate primer from the previous step. To do this we use a random forest regressor model trained from prior experimentation (discussed in Amplification efficacy model). In short, this non-linear regression model is trained on plasmid experiments conducted using sets of a single primer and a single plasmid genome. The goal of this regressor is to predict amplification efficacy from various properties of the primer, including computed thermodynamically-principled features estimating a primer's binding affinity for the target genome (see Random forest features). After predicting an on-target amplification value for each of the candidate primers from the previous step, the primers are filtered out according to the minimum predicted on-target amplification threshold (`min_amp_pred`) parameter, which by default is set to 5. This step significantly reduces search computation by weeding out low-amplification primers.

Step 4. Primer set search and evaluation. Using the filtered list of primers from the previous step, SOAPswga searches for primer sets using a machine-learning guided scoring function and a breadth-first, greedy approach. At a high level, `max_sets` number of top primer sets are built in parallel, primer by primer, by adding primers which increase the evaluation scores the most. More specifically, we run the following algorithm.

Algorithm 1: Primer Set Search

```
Input: primer_list (list of candidate primers)
       max_sets (maximum number of sets to explore at each step)
       amp_efficacy_scores (amplification efficacy scores predicted from random
 forest model)
       drop_indices (defines which iterations are dropout layers)
Output: top_sets
top_sets ← RandomInitialStart (primer_list, amp_efficacy_scores, max_sets);
top_scores ← Evaluate (top_sets);
curr_sets = ← [];
curr_scores ← [];
for i = 1 to max_iterations do
    for top_set in top_sets do
        for primer in S do
            if Compatible (top_set ∪ [primer]) then
                new_set ← [top_set ∪ [primer]];
                score ← Evaluate (new_set);
                curr_scores ← curr_scores ∪ [score];
                curr_sets ← curr_sets ∪ new_set;
            end
        end
    end
    if i is in drop_indices then
        top_sets, top_scores = DropOut (curr_scores, curr_sets, max_sets)
    end
    else
        top_sets, top_scores = ChooseMaxSets (curr_scores, curr_sets, max_sets)
    end
end
return top_sets
```

The function `RandomInitialStart` randomly chooses the first primer in each of the `max_sets` number of primer sets where probabilities are the normalized amplification efficacy scores from the previous step. This allows for randomization of the search initialization and permits better exploration of the search space. The function `Compatible` checks that no two primers have subsequences longer than `default_max_dimer_bp` that are complementary. This is done in an effort to reduce the risk of primers binding with each other rather than the target genome. The function `DropOut` allows each of the highest-scoring primer sets to drop one primer by taking the subset of size one less than the current set with the highest evaluation score. This is particularly useful, for example, if a primer set has an evaluation score much higher than the primer set excluding the initially chosen primer. "Dropping" a primer also allows for the possibility of adding a primer that would otherwise be barred because of its risk in forming dimers with the dropped primer. `ChooseMaxSets` simply chooses up to `max_sets` number of best sets based on evaluation scores. `Evaluate` evaluates the primer set using the following equation: Score = β_0 + β_1 `freq_ratio` + β_2 `mean_gap_ratio` + β_3 `coverage_ratio` + β_4 `on_gap_gini` + β_5 `off_gap_gini` where Table 4.1 contains the coefficient values, and the score is a prediction of the percent coverage ($1\times$) per base pair sequenced. This step can be re-run multiple times and includes the option of withholding primers too frequently used in previous primer sets until after the dropout layer.

Primer set evaluation model

Training data. Here we discuss how we developed the function for scoring primer sets given a target and off-target genome. In order to learn this function, we used the 46 sets of SWGA data for *M. tuberculosis*/*H. sapiens* (3-6 trials each) from [18]. From the bam files, we computed the percent genome coverage at $1\times$, scaled by the sequencing effort (number of base pairs sequenced), which serves as the dependent variable to be predicted. This metric is also discussed in [18].

Feature selection. The model uses five variables, which are all normalized appropriately by the genome lengths so as to generalize to other genomes. The variables and their respective coefficients values are described in Table 4.1. The variables `freq_ratio` and `mean_gap_ratio` are summary statistics that Clarke et al. [18] found to either or both be useful for assessing amplification and should be large and small, respectively (equivalent to positive and negative regression coefficient values). To understand why we included the remaining features, we note that `mean_gap_ratio` actually measures the average gap sizes in the genome, using positions on both the forward and reverse strand where the computation is indifferent to which strand a binding position lies. In theory, however, each binding position on the forward strand should have at least one position downstream on the reverse strand in close proximity (see Figure 4.2). One way to measure this is to sum the number of positions that are within 70 kbp distance (the observed maximum length of the synthesized DNA chain [9]) for each binding site on the forward strand. Normalizing by the total genome

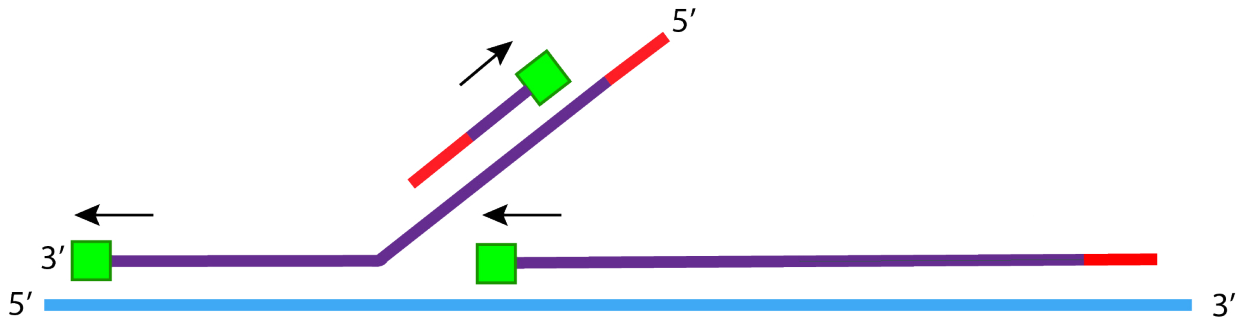


Figure 4.2: Simple depiction of multiple displacement amplification. The red lines indicate a primer while the green blocks indicate a ϕ29 enzyme. The blue line is the original 5' to 3' template strand and the purple indicates DNA polymerized by ϕ29. For exponential amplification, each primer binding site on the 5' to 3' strand should have a binding site in close proximity (within 70 kbp distance–the observed maximum length of the synthesized DNA chain ([9])) on the 3' to 5' strand. In this way, primers on opposing strands work together to alternately amplify the region in the forward and reverse direction.

length gives a rough approximation of coverage, and the ratio of this value computed using the off-target to that computed using the on-target is referred to as the `coverage_ratio`, which we would like to be as small as possible, i.e., have a negative regression coefficient value. Additionally, for evenness purposes, we want binding sites to be evenly distributed on each individual strand—not necessarily on the genome, as a whole. To measure this, we use `on_gap_gini` and `off_gap_gini` which averages across strands the Gini index of distances between binding sites.

Aside from the described features, we tried a number of other features, which included variables like the number of nucleotides between on-target binding sites that exceed 70 kbp as well as entropy and generalized entropy of the on-target binding site distribution. For the off-target binding site distribution, we experimented with summary statistics such as kurtosis, skewness, bimodality, and variance.

Model selection. We tested a number of different models built from various feature sets, and normally model error would be our only guiding intuition for which model to select. However, in our unique case we were able to leverage domain knowledge—that is, for most of these variables, we have a strong intuition for what the sign of the coefficient should be. For example, we know that the model should be encouraging the total number of on-target binding sites and penalizing the number of off-target binding sites, meaning the signs should be positive and negative respectively. After filtering out the nonsensical models, we performed a 10-fold cross validation and selected the best model according to cross-validation error.

Given so few data points, the risk of overfitting and not being able to generalize to

Variable name	Variable description	Coef.	Coef. value
intercept	Intercept	$\hat{\beta}_0$	-3.14×10^{-15}
freq_ratio	Ratio of the binding site rate in the on-target to off-target.	$\hat{\beta}_1$	0.321
mean_gap_ratio	Ratio of the mean distance between binding sites of the on-target to off-target genome, aggregating across strands.	$\hat{\beta}_2$	-0.0368
coverage_ratio	Ratio of the coverage approximation of the on-target to that of the off-target.	$\hat{\beta}_3$	-0.0318
on_gap_gini	Mean gini index of on-target binding site gap sizes, averaging across strands.	$\hat{\beta}_4$	-0.0131
off_gap_gini	Mean gini index of off-target binding site gap sizes, averaging across strands.	$\hat{\beta}_5$	0.281

Table 4.1: Ridge regression variable descriptions and respective coefficient values for primer set evaluation. SWGA experiments from Clarke et al. [18] were used as training data.

new sets is high. Consequently, we utilize ridge regression which uses regularization to upper bound the complexity and to reduce the variance of the model, without a substantial increase in bias. Because variables on a larger scale are unfairly penalized by regularization more so than others, we use standardized features for training by subtracting the data by the respective means and dividing by the respective standard deviations.

Amplification efficacy model

Here we discuss the development of the random forest regressor used to predict the amplification efficacy of individual primers in step 3 of the pipeline. In the SWGA setting, multiple primers are usually used on large genomes, making it challenging to isolate the efficacy of a single primer. Thus, we performed a series of rolling circle amplification (RCA) [3, 25] experiments of circular plasmid DNA and a single primer using Φ29 DNA polymerase. From this data, which was iteratively collected in an active learning fashion, we trained a model using various properties of the primer sequence and binding affinities for the target genome, estimated by the thermodynamic nearest-neighbor DNA model [99].

Plasmid experiments. Our RCA reactions were designed to (*i*) use a primer with one exact binding site to exponentially amplify target template plasmid DNA [3] and (*ii*) use the same single primer to assess the degree of amplification when there were no exact binding sites on the off-target plasmid DNA. We used plasmids because they allowed for long-range amplification similar to what we would expect with most genomes.

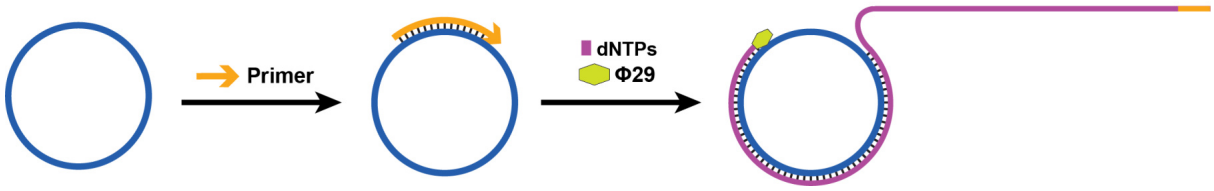


Figure 4.3: Illustration of singly-primed rolling circle amplification.

We obtained plasmid DNA from Addgene (Watertown, MA) and chose two plasmids of equivalent length (approximately 6kb). pcDNA3-EGFP was provided to Addgene by Doug Golenbock (Addgene plasmid #13031; http://n2t.net/addgene:13031; RRID:Addgene_13031) and pLTR-RD114A [127] was provided to Addgene by Jakob Reiser (Addgene plasmid #17576; http://n2t.net/addgene:17576; RRID:Addgene_17576). We isolated plasmid DNA from liquid cultures of bacteria containing each plasmid using the QIAGEN Plasmid Mini Kit (Qiagen, Valencia, CA). We verified the identity of the purified plasmids using a restriction digest analysis. We digested our plasmids using EcoRI and MscI individually and in combination (Thermo Fisher Scientific, Waltham, MA), both of which have different numbers of cut sites on each plasmid. We confirmed the size and conformation of both undigested and digested plasmids using gel electrophoresis.

In a set of experimentation, 96 primers were chosen, and each primer was used in two reactions, one with the target plasmid and one with the off-target plasmid. In order ensure that primers would not be degraded by the exonuclease activity of Φ29 DNA polymerase, we added two 3′-terminal phosphorothioate modifications to every primer We ran the amplification reactions using a modified version of previously published methods for singly-primed RCA [40] (see Figure 4.3). We mixed 5 ng of template plasmid DNA with 5 µL of 10X Reaction Buffer (New England Biolabs, Ipswich, MA), 1 µL of 10 mg/mL BSA (New England Biolabs), 1 µL 10 mM dNTPs (New England Biolabs), 2.5 µL of 10 µM primer (final reaction concentration of 0.5 µM of primer), and molecular grade water to 49 µL. We denatured these mixtures at 95°C for 3 minutes and rapidly cooled them to 4°C and placed them on ice. We amplified the denatured plasmid DNA using 1µL of Φ29 DNA polymerase (10 U/µL, New England Biolabs) with a one-hour ramp-down step from 35-30°C [55], followed by a 16-hour amplification step at 30°C, and a denature step for 15 minutes at 65°C. Following amplification, we measured the resulting concentration of DNA in each reaction using the Qubit^{TM} dsDNA HS Assay Kit.

An active learning approach. Active learning, a type of iterative supervised machine learning, was used to maximize information gain per each experiment. In other words, we purposely did successive rounds of experimentation, improving our model with each round and querying primer/plasmid combinations with specific goals in mind. In total, we designed three rounds of experiments, each of 96 primers, except the initial experimentation round

which had 204, designed to differentially target two different plasmids as explained in the previous section.

For Round 1, we first used a previously published SWGA Perl script [55] to generate a list of 'uniquely targeted' primers for each plasmid (i.e., a primer with a binding site that occurred once in the 'target' plasmid sequence and zero times in the 'off-target' plasmid sequence). We then tried to pick primers based on a discretization of the search space in terms of sequence length, melting temperature, GC content, the proportion of different nucleotides, longest G repeat, longest C repeat, longest A repeat, and longest T repeat. Amplification values for Rounds 2 and 3 primers were predicted based on the random forest regressor to be discussed in the following sections.

Random forest features. Before we discuss model selection and training, we first describe the model features. To start, we chose 22 primer attributes for which we had *a priori* evidence may impact the efficacy of a primer to accurately bind to template DNA and to promote amplification under standard PCR or RCA conditions [26]. See Table 4.3 for a list of these features and their feature importances. We quantified and characterized all candidate primers tested in this study for these 22 primer attributes. However, these 22 attributes such as melting temperature are not specific to the target genome. While computing the exact matches of a primer in a genome is a good heuristic for a primer's ability to amplify, using exact matches is a coarse metric. Therefore, we included features which capture the thermodynamic likelihood of the primer binding along the genome using a unified thermodynamic nearest-neighbor DNA model [99]. This model is widely used to calculate primer melting temperature in a wide variety of available primer design tools [98]. Empirical thermodynamic parameters (ΔG°_T) are available for most primer binding and single mismatch scenarios and include thermodynamic data for initiation, symmetry correction, all Watson-Crick pairs, Terminal 'AT' penalties, and internal mismatches [99]. Empirical thermodynamic data for terminal mismatches are not publicly available [99, 98], and were not incorporated into our predictive model.

Measuring binding affinities is not sufficient, however, since all genomes are not the same length. Consequently, we need a transformation of the features such that the resulting features are invariant to the genome length. In order to apply this thermodynamic model to produce features that can be used for any genome and any primer, we do the following for each primer and each genome:

1. Compute the ΔG°_T values for each position in the genome according to [99].
2. Compute a histogram of the list of $\Delta G^\circ_T < 3$ with bins ranging from -20 to 3.
3. Scale the counts by the length of the genome.

Step 1 produces a list of ΔG°_T at every position, which is a smoother metric for a primer's propensity for binding along the genome than the number of exact matches. Step 2 is necessary for converting this list of values into a fixed number of features, which is necessary

for building a model. Lastly, in Step 3 we normalize by genome length because we would like our model to be applicable to different sized genomes. In the end, we have features that capture the percentage of positions that have a particular binding propensity with the primer.

Model selection and hyperparameter optimization. We tested a number of different regressors: linear, logistic, random forest, gradient boosting, and support vector machine. We found that non-linear regressors did much better than linear models, and in particular the random forest regressor from the `sklearn.ensemble.randomforestregressor` module in python performed the best. We also did a hyperparameter search and found that the optimal parameters were `n_estimators`=1500, `min_samples_split`=10, `min_samples_leaf`=4, `max_depth`=50, `bootstrap`=False.

4.3 Results

Plasmid experiments

As touched upon in our discussion of an active learning approach, we conducted experiments in 3 stages, so as to maximize information gained from each experiment. In the initial round 1, there was no prior data for this type of experiment, so the goal in this iteration was to maximize our exploration of the search space and test 204 primers with varying length, GC content, and molarity. In Round 2, we used data collected from Round 1 to train a random forest regression model to accurately predict the amplification of 96 primers. Finally, in Round 3, we focused on predicting 96 high-amplification primers.

Non-linear model is able to predict high-amplification primers after two rounds. In the initial Round 1 the primers were selected near randomly, but as apparent in Figure 4.4A, which shows plots of the target amplification per round, Round 1 did not have many high-amplification primers. In Round 2, we used data collected from Round 1 to train a random forest regression model to predict primers with poor, mediocre, and high amplification. Figure 4.4A additionally demonstrates that in Round 2 many of the primers also did not have high amplification, likely because Round 1 data did not have many high-amplification primers, which made extrapolation to the high-amplification regime difficult. After two rounds, however, the non-linear model was able to learn from the data sufficiently enough to predict many more high-amplification primers (Figure 4.4A).

Low-amplification regime has lower variance and can be predicted more accurately. For Round 3, we retrained the model on data collected from both Round 1 and 2, focusing on predicting high-amplification primers. Though many more high-amplification primers were predicted in Round 3 than in the first two rounds (Figure 4.4A), the error of the predictions measured in terms of the mean squared error increased (Figure 4.5A). This is

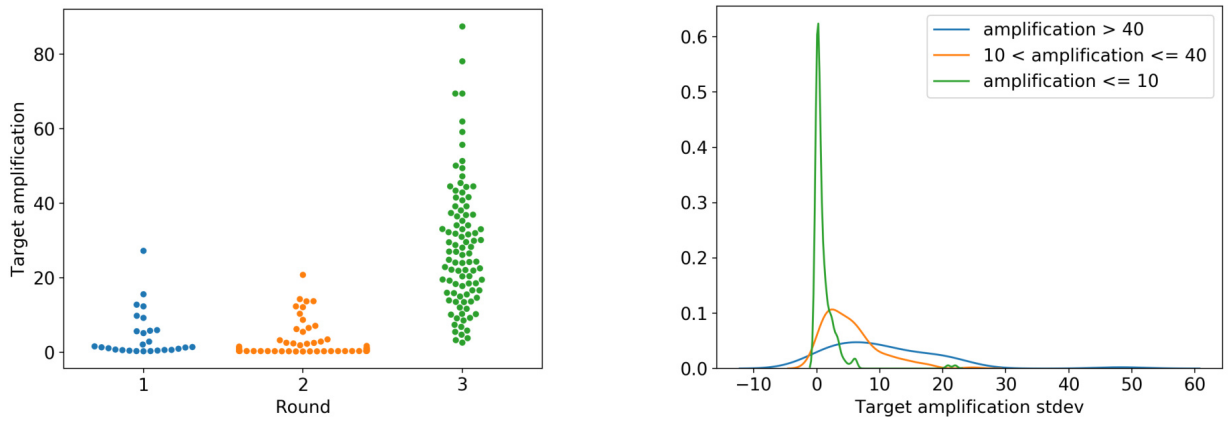


Figure 4.4: Target amplification per round and replicate standard deviation per amplification regime. (A) Categorical dot plots of target amplification in each round of experiments, controlling for molarity = 0.5 μ. Points are adjusted along the axis so that they do not overlap. (B) Standard deviation of target amplification among three trials per experiment for all the data. The primers in the high-amplification regime exhibit a larger standard deviation.

likely due to the large variance in terms of experiment replicates (Figure 4.4B), which makes accurate prediction of high-amplification primers a much more difficult problem than the prediction of poor amplification primers given the increased noise. However, overall it seems that the model has a high accuracy rate in terms of predicting poor amplification primers, i.e., primers with true amplification scores less than 10 (Figure 4.5)). In other words, if the regression predicts a primer will have amplification less than 10, the true amplification is often also less than 10. For this reason, we reframe our regression as a classification problem, using this model as a basis for filtering out low-amplification primers in Step 3 of the pipeline. Table 4.2 reports the average percent of primers that were incorrectly filtered out over 100 iterations using the random forest regressor trained on 75% of the data and tested on the mutually exclusive 25% of the dataset. Based on the results, a default threshold of 5 was chosen, given its small error rate and substantial 26% filtration rate, which can be hugely beneficial for reducing the time complexity of the program. The default value of 5 is customizable according to the parameter `min_amp_pred`.

Feature importances indicate significance in binding affinities. Having trained this random forest model, a natural way to understand the significance of each feature is to investigate the feature importances, which are computed from the variance reduction from each split in each tree (see Table 4.3). The third column indicates that the binding affinities are informative to the model while the last column of Table 4.3 aggregates the feature importances for subsets of features. The binding affinities account for 16% of the feature

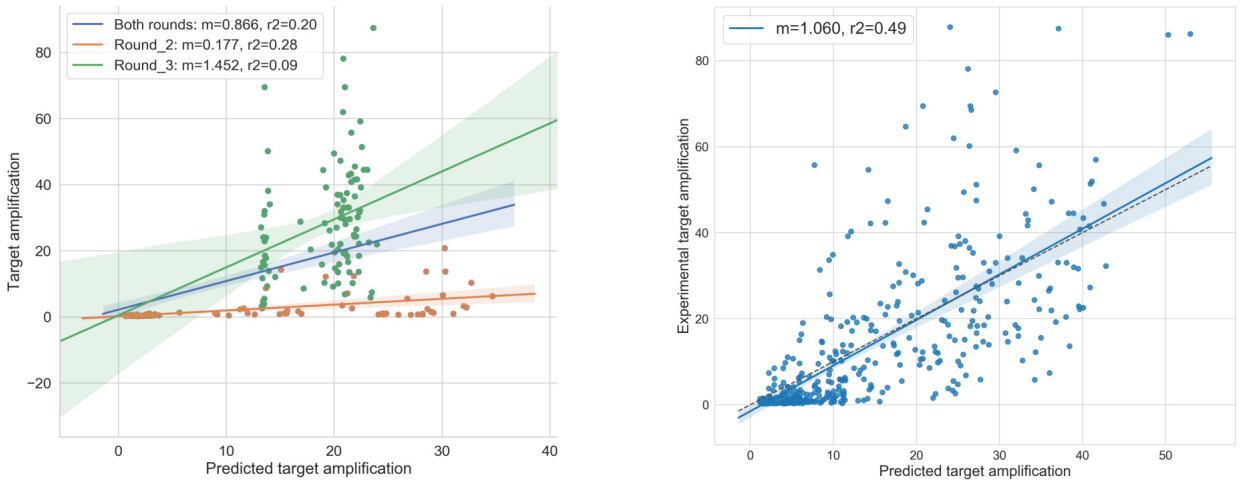


Figure 4.5: Predicted versus true target amplification for *Mycobacterium tuberculosis*. The left displays regression lines based on the random forest regression trained only on data prior to that round. The right displays the regression fit based on the model trained on all the data. (A) Predicted versus experimental amplification for Rounds 2 and 3. Round 3 has much higher margin of error because of higher variance in experimental trials with high-amplification primers. Round 1 primers were not predicted. (B) Predicted versus experimental amplification value after retraining the random forest regression on all the data (396 datapoints), essentially showing overall training error which is smallest in the low amplification (< 10) regime.

Threshold	Percent incorrect	Percent filtered
2	0.04	6.51
5	1.62	26.5
10	3.76	47.1
15	4.42	58.9
20	5.44	66.6

Table 4.2: Summary statistics of various thresholds for filtering out low-amplification primers after scoring with the random forest regressor. For each threshold, we report the percent incorrect (number of primers incorrectly filtered out—i.e., the predicted amplification is less than the threshold but the true amplification is greater than the threshold) and percentage of primers filtered out. Percentages were computed from averages of 100 iterations, training on 75% of the data and testing on a hold out set of 25%.

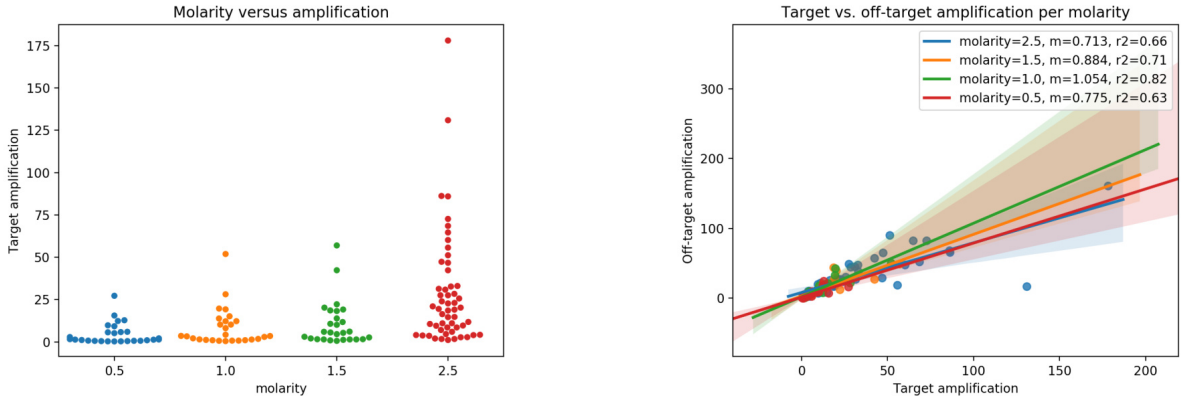


Figure 4.6: Target and off-target amplification of 13 primers at 4 different molarities. (A) Target amplification per molarities 2.5, 1.5, 1.0, and 0.5 µ using a set of 13 primers. (B) Target vs off-target amplification according to molarities 2.5, 1.5, 1.0, and 0.5 µ.

importances, which could be because they capture more information than exact matches, and in particular, capture information about binding sites which may have a few mismatches but may still be a binding site. Perhaps unsurprisingly, features involving GC content or the number/proportion of G's or C's are the most important subset of features. In fact, selecting a random decision tree from the random forest model will often use a decision rule suggesting that greater numbers of G's or C's will contribute to higher amplification power. This is in agreement with the thermodynamic nearest-neighbor DNA model because the more G's and C's in a primer, the more negative the ΔG°_{T} values, according to the model in [99]. Additionally, many of the decision rules involving molarity suggest that amplification increases with molarity. This is in line with the findings in the proceeding paragraph.

Primer concentration affects amplification. In Round 1, we experimented with 4 different molarities of the same 13 primers—the four molarities being 2.5, 1.5, 1.0, and 0.5 µ. As one might expect from the principles of chemical equilibrium, an increase in the concentration of primers increases the amount of amplification product (Figure 4.6A). It does not, however, seem to suggest that molarity affects the target to off-target amplification ratio (Figure 4.6B).

M. tuberculosis and *H. sapiens* SWGA experiments

Using the pipeline, we experimented with nine primer sets (A1-A9) targeting the genome *Mycobacterium tuberculosis* where the off-target genome is *Homo sapiens*. For these sets of primers (which we will refer to as "Round A") we used a simpler evaluation function, which does not incorporate `coverage_ratio` and measures the Gini index from aggregating

Subset Description	Feature Description	Feature Imp.	Subset Feature Imp.
sequence length	sequence length	2.56	2.6
molarity	molarity	10.2	10.2
melting temperature	melting temperature	6.33	6.3
G/C content features	number of G's	11.8	27.9
	proportion of G's	2.68	
	number of C's	2.76	
	proportion of C's	6.54	
	GC content	4.08	
A/T content features	number of A's	1.16	6.42
	proportion of A's	1.25	
	number of T's	1.03	
	proportion of T's	2.98	
repeat features	longest A repeat	0.43	19.2
	longest T repeat	0.61	
	longest G repeat	4.10	
	longest C repeat	3.51	
	AA repeat number	0.33	
	CC repeat number	2.37	
	TT repeat number	0.46	
	GG repeat number	7.30	
last five bases features	GC.clamp	1.92	10.2
	first base from 3' end	1.08	
	second base from 3' end	2.48	
	third base from 3' end	1.32	
	fourth base from 3' end	1.41	
	fifth base from 3' end	1.96	
binding affinities	-20 to -12	0.87	18.1
	-12 to -6	1.16	
	-6 to 0	5.47	
	0 to 3	10.5	

Table 4.3: Percent feature importances of the random forest regressor model. We include aggregated feature importances (computed by summing the individual feature importances in the subset) as an additional means of interpretation. The 27 binding affinity features were aggregated for brevity and clarity.

positions on both the forward and reverse strands. We will refer to this evaluation function as "Function A" and the function first described in Section 4.2 as "Function B". Moreover, instead of taking the top 9 primers from the pipeline, we tried to diversify the sets by limiting too much intersection between sets. Lastly, we included one primer set (A5) which had a low predicted evaluation score in order to have a more informative benchmark of our evaluation function.

Computational costs reduced from weeks to minutes. Creating files which store information on all 6-mers to 12-mers for *Mycobacterium tuberculosis* and *Homo sapiens* takes a few hours, but needs to be done only once. Using these precomputed files and a 2013 MacBook Pro with a 2 GHz Intel Core i7 with 16 GB of memory, where the `max_sets` is set to 5 (the algorithm searches for 5 sets in parallel) and `max_primers` is 200 (optimization is done over 200 primers), the pipeline runs in roughly 58 minutes. This is a huge improvement from the `swga` which runs for an estimated two weeks. While it is difficult to benchmark the two programs given that it would take weeks to do so, it is clear that multi-processing in many aspects of the pipeline and the change to an efficient rather than exhaustive search algorithm greatly improves the speed. Additionally, we have used new file formats like h5py, which allows for O(1) read access to the binding positions of a particular primer, and added new data structures in the search algorithm, which stores the score for every primer set evaluated, barring unnecessary recalculation for future iterations. The way in which the current pipeline computes Gini index is also modified; it is no longer an approximation (which can be unstable in small samples) but an efficient calculation of the exact Gini index using an alternate expression of the equation and taking advantage of efficiencies in array computations in the Python library `numpy`.

Experimental values far exceed those of previous experiments and predicted scores. Figure 4.7 shows the side-by-side bar plots of Round A primer sets and the primer sets from Clarke et al. [18] evaluated for two metrics. The first is "% Reads Mapped" which measures the percentage of reads which mapped to *Mycobacterium tuberculosis*. The second metric is "Percent coverage per bp sequenced" which is the percent of the genome with $1\times$ coverage, normalized by the total base pairs sequenced and scaled by 10^7 for readability. Primer sets with an "A" in the prefix refer to those evaluated by our pipeline and far exceed the results from Clarke et al. [18]. However, looking at Mtb6 and Mtb9, which we ran alongside the Round A primer sets, we can see that some of these differences are likely due to a difference in experimental setup, the lack of replicates, and/or the lower sequencing depth at which Round A primers were sequenced. Indeed, the average total number of reads sequenced for the Clarke sets was 1.5 million whereas for the Round A primer sets it was 21,000. We intend to rerun these primer sets at comparable sequencing depth, given that Figure 4.7 displays promise. Compared to Mtb6 and Mtb9, many of the Round A primers that we predicted would do reasonably well performed on par.

In addition to surpassing the results reported in Clarke et al. [18], the results from Round A also greatly surpassed predictions. Figure 4.8 summarizes these predicted values alongside the second metric, percent coverage per bp sequenced. Predictions were done using both evaluation functions for Round A primers as well as Mtb6 and Mtb9 which were two of the most successful primer sets in Clarke et al. [18]. Overall it seems that the two functions perform similarly in terms of ranking the Round A sets, but whereas Function A scored A7, A8, and A9 on par with Mtb6 and Mtb9, Function B puts them much lower and consequently would not have been selected. In fact, if we count the number of sets which have predicted values greater than Mtb6 but experimental values less than Mtb6 (number of false positives) there are 3 using Function A and 0 using Function B. In this sense, Function B seems preferable because there is greater confidence in the sets scoring greater than the baseline Mtb6. While the predicted values certainly fall short of the true values of A1, A2, A3, A4, and A6, it's understandable when looking at Figure 4.9, which plots the experimental value against the predicted values. The black dots plotted indicate the training set, while the blue/orange dots are the Round A primer sets. It is clear that for the function to correctly predict the outliers (particularly A1 and A4) requires extrapolation, for which there is likely not enough data. It also remains to be seen if the outliers are an artifact of low sequencing depth.

4.4 Discussion

SWGA is a highly complex set optimization problem involving scalability challenges and limited prior data, which is often also noisy. In *Mycobacterium tuberculosis* alone there are over five million candidate primers that must be filtered into a reasonable working set. On the practicality side, one of the immediate challenges in implementing a solution for SWGA is designing data structures and files which efficiently store the information needed for evaluation without having to make expensive searches of a genome for potential binding locations. These challenges are exacerbated by large genomes, particularly the human genome which is over three billion base pairs. Probably the greatest challenge is identifying a way to search and evaluate primer sets, the former which demands combinatorial optimization techniques and the latter which demands understanding prior data. Understanding and learning from prior data has its own challenges, due to the limited availability of data, often high variability in replicates, and indirect inference of the polymerization process via short read data.

SOAPswga offers a number of novel improvements. Most obviously is the difference in speed due to the implementation of parallelization, caching, and branch-and-bound techniques as well as storing data in efficient formats like h5py. Additionally, computations of Gini index are no longer an approximation and take advantage of efficiencies in array operations. In step 1, we use `jellyfish` [69] which is much faster than DSK [93] used in `swga`. In step 2 we have added a number of new filters recommended for good primer design and prevention of self-dimers. In step 3 we have added a novel machine-learning model for scoring individual primers based on amplification capability, which uses a unique set of

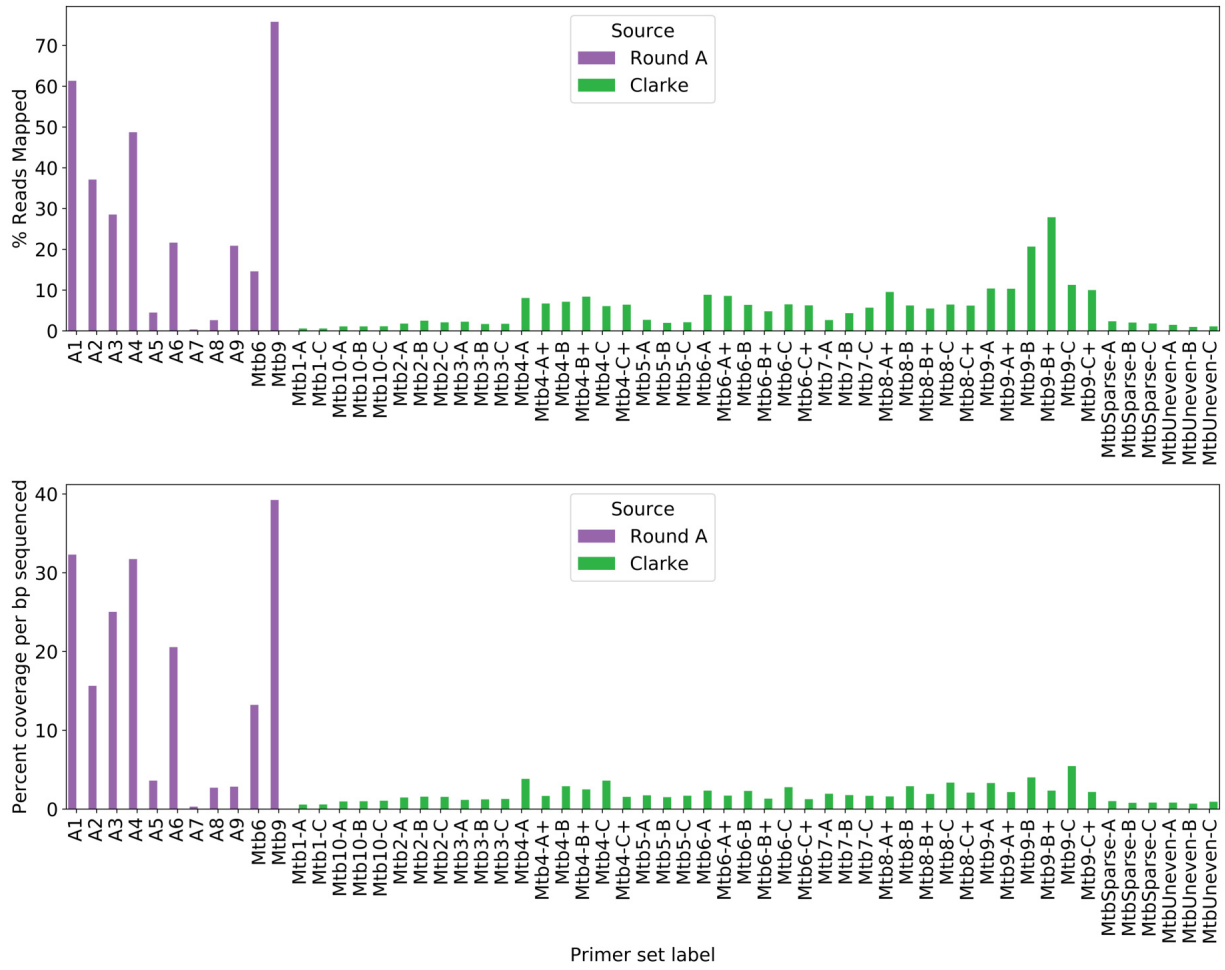


Figure 4.7: Bar plots of *Mycobacterium tuberculosis* results for both the primer sets from Clarke et al. [18] and from using evaluation function Function A in SOAPswga—primer sets which we term "Round A". The top bar plot measures all primer sets according to the percentage of reads which mapped to *Mycobacterium tuberculosis* and the second metric which measures the percent genome coverage at 1×, normalized by the total number of base pairs sequenced. The Clarke primer sets have a dash and a suffix appended to them, which indicates which experiment replicate it is. For the Round A primer sets, only one replicate was done. Mtb6 and Mtb9 from Clarke et al. [18] were also done alongside Round A primer sets for benchmarking purposes.

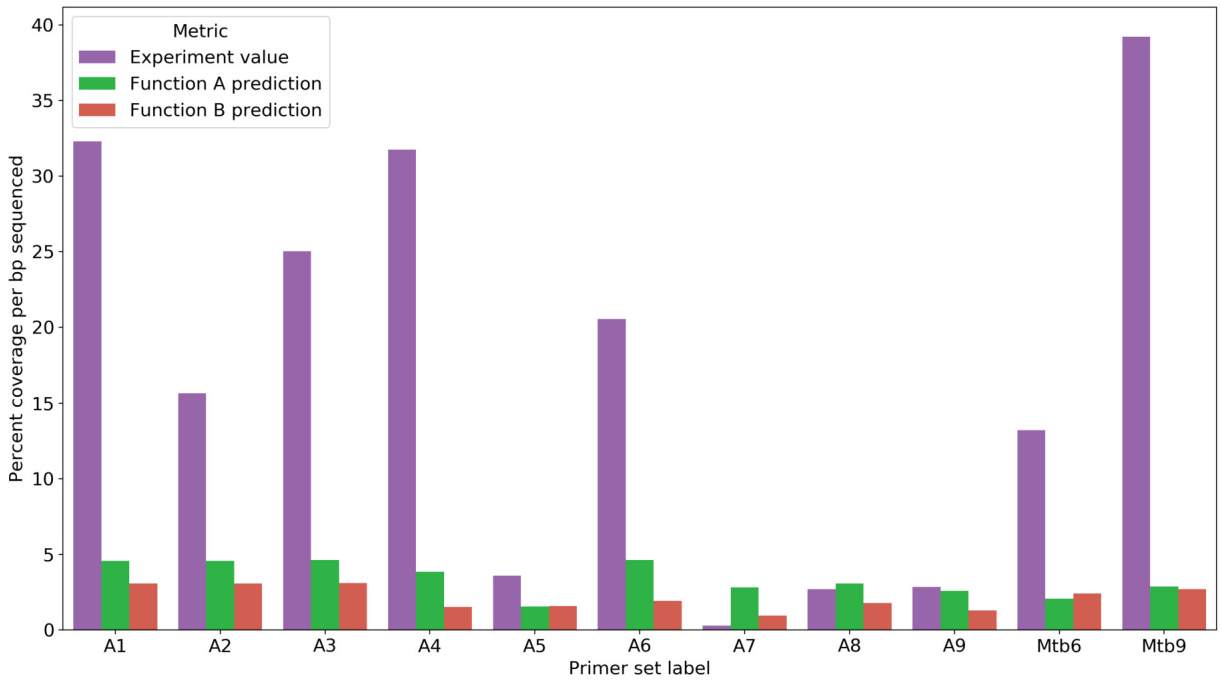


Figure 4.8: Bar plots of predicted and experimental values for percent genome coverage at 1× (normalized by the total number of base pairs sequenced) of Round A primers as well as Mtb6 and Mtb9, which were two of the most successful primer sets in Clarke et al. [18]. Predictions were done using Function A and Function B, which is described in the Methods section. Mtb6 and Mtb9 were repeated alongside the Round A primers for benchmarking purposes.

thermodynamically-principled binding affinities. Lastly, while predecessor software to this pipeline made great strides in implementing methods for SWGA, optimization is no longer done by hand or via exhaustive search. In step 4 we implement branch-and-bound techniques and utilize evaluation functions learned from prior data using machine learning. To better explore the search space, we also employ randomization and a novel concept of dropouts.

In our results, we demonstrate the ability of the random forest model to meaningfully learn from the data after two rounds of data, the accuracy of the random forest model in terms of weeding out low-amplification primers, and the utility of our novel thermodynamically-principled binding affinities. Moreover, our results demonstrate a large reduction in terms of time complexity, and the SWGA results exhibit promise in terms of performance. For future work, in addition to replicating Round A primer sets, we intend to test a new set of primer sets, which we will call "Round B". Round B sets were chosen using Function B, and ensuring no two primers sets had an intersection of 4 or more. While we cannot evaluate Round B in terms of experimental results, we benchmark the sets according to its

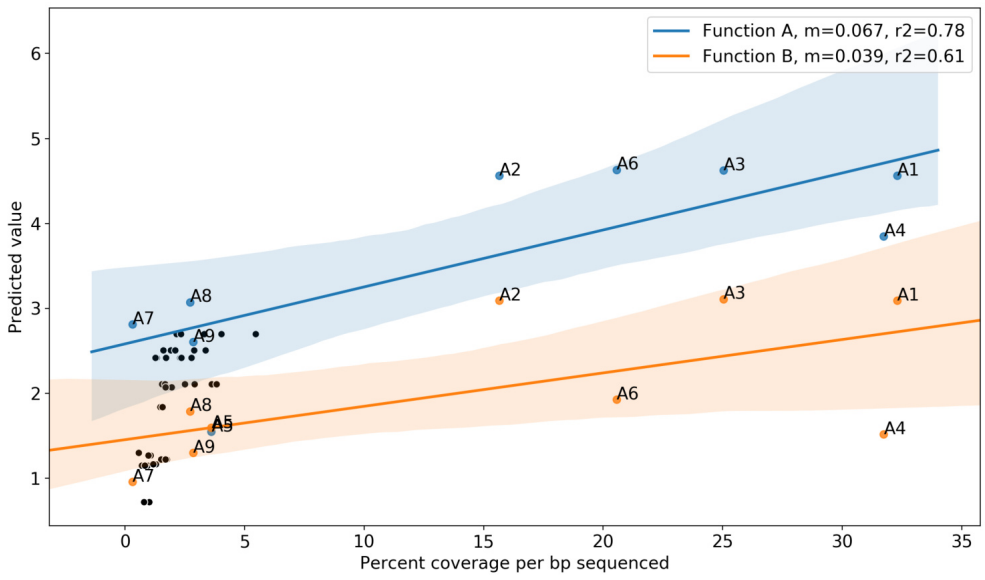


Figure 4.9: Experimental values plotted against predicted values for percent genome coverage at 1×, normalized by the total number of base pairs sequenced according to Function A and B. Predictions according to Function A are shown in orange, and predictions according to Function B are shown in blue. Black dots correspond to experimental and predicted values according to Function B of the primer sets from Clarke et al. [18]. The predicted values greatly underestimate the experimental values likely because of the low sequencing depth of Round A results and/or the lack of training data with values in the high range. The largest experimental value of the Round A primers is roughly 7 times that of the Clarke primer sets, and extrapolation to that extent, would have seemed unreasonable.

predicted scores and the average mean gap distance in the target genome, which Clarke et al. [18] found important and negatively correlated with amplification. The average mean target distances of the Round B primer sets is roughly a fourth of that of the Clarke primer sets. In addition, the evaluation scores according to both Function A and Function B are well above those of the previous sets.

This pipeline could be used in a number of contexts where primer design is critical such as in generating forensic DNA profiles, detecting pathogens during infection, or species identification in metagenomics. In many such scenarios, it would be undesirable to have primers that also favor other genomes that are present in the same sample. In some contexts, there may be separate portions of the same genome which may serve as the "target" and "off-target" genome. For example, base editors like CRISPR have seen a number of undesirable off-target mutations [45, 130], which have the potential to be incredibly deleterious. The design of gRNA sequences would be a natural application of the same principles used to design primer sets for SWGA.

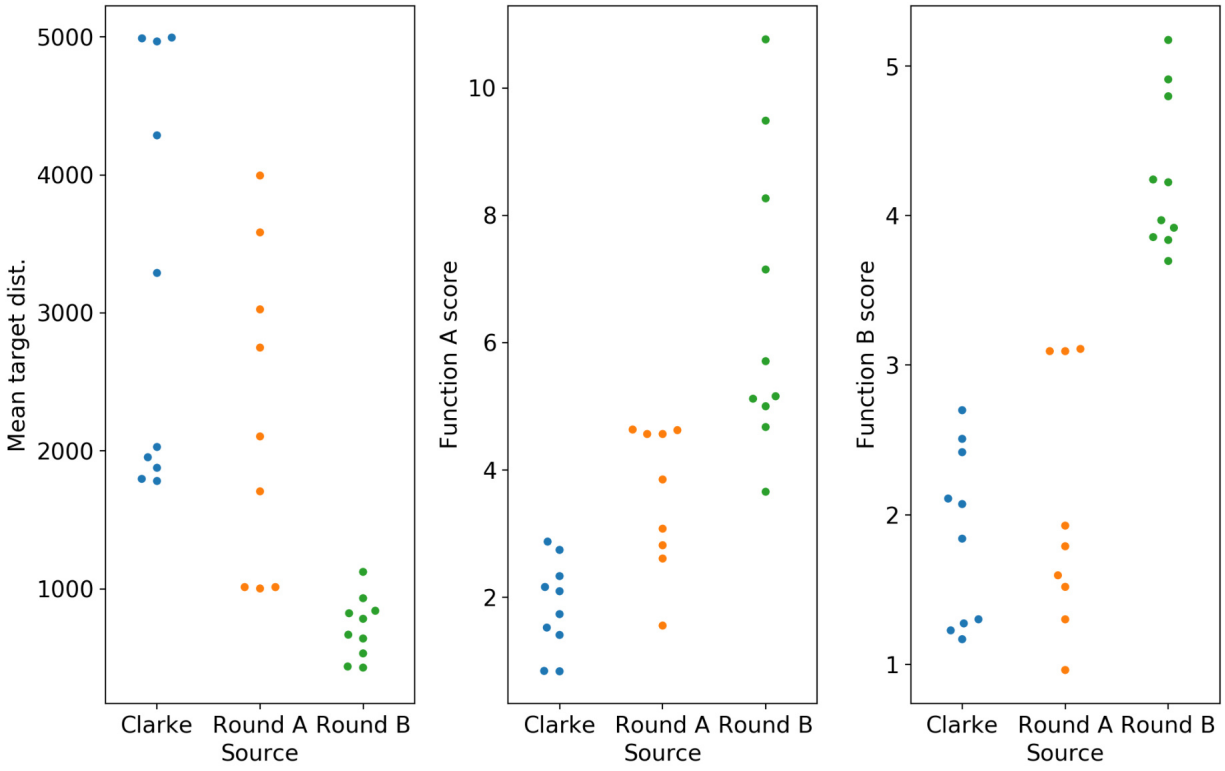


Figure 4.10: A comparison of all three primer sets (those from Clarke et al. [18], Round A, and a proposed Round B) according to the mean target distance, and scores according to Function A and B. The average mean target distance for the Clarke, Round A, and Round B primers is roughly 3200, 2240, and 720, respectively.

Chapter 5

Conclusions

In this dissertation, we take a close look at two distinct problems in the subfield of computational genomics. In the first two chapters, we present a customized genotyping pipeline that produces an accurate characterization of the IGHV and TRBV loci without the need to reconstruct the complete sequence. The IGHV and TRBV loci are genomic regions critical for the adaptive immune system and hence of great medical relevance. Because these loci are comprised of numerous duplicated genes that are highly similar and harbor substantial copy number variation, cataloging and analyzing genetic variation in these gene families has remained challenging thus far. We have also applied our method to the first study of both the IGHV and TBRV loci in a globally diverse sample of humans, in addition to simulated data and a family composing three generations. Our work provides a quantitative analysis of genetic variation relevant to the fields of population genetics and medicine, as well as to the greater understanding of the long-term evolution of these gene families.

In the penultimate chapter, we turn to a separate problem in sequencing—selective sequencing of a target genome in a heterogeneous sample. Given the correct primer set, Φ29 multi-displacement amplification technology is the most inexpensive, flexible, and shareable culture-free technology for amplifying a specific genome. While previous methods utilize exhaustive search based on heuristic functions, SOAPswga is an efficient pipeline integrating machine learning and optimization principles for proposing primer sets. We present an application of this pipeline to the target genome *Mycobacterium tuberculosis* in a sample of human blood, benchmarking against previous work. This work is not necessarily limited to the context of SWGA, but can be extended to other contexts involving targeted genomic regions and provides insight into individual primer efficacy relevant to primer design in general.

As more reference sequences become available and as sequencing technology improves, we expect the power and precision of these tools to improve. Assembly of the IGHV and TRBV genes, for instance, would become more reliable given longer and less error-prone reads. Additionally, with longer reads, the endpoints and distribution of DNA strands polymerized by $\phi 29$ could be better estimated, which would help in understanding the exact behavior of multiple displacement amplification. These problem settings will also undoubtedly benefit from the growth in available whole-genome sequencing data and computational power in an

age where sequencing is cheaper than ever before, and GPUs are becoming more powerful and increasingly commonplace.

Bibliography

[1] A. Abyzov et al. "An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing". In: *Genome Res* 21 (2011), pp. 974–984.

[2] Bruce Alberts et al. "The generation of antibody diversity". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[3] M. Monsur Ali et al. "Rolling circle amplification: a versatile tool for chemical biology, materials science and medicine". In: *Chemical Society Reviews* 43.10 (2014), pp. 3324–3341. ISSN: 0306-0012.

[4] C. Alkan et al. "Personalized copy number and segmental duplication maps using next-generation sequencing". In: *Nat Genet* 41 (2009), pp. 1061–1067.

[5] Hatim T Allawi and John SantaLucia. "Thermodynamics and NMR of internal G.T mismatches in DNA". In: *Biochemistry* 36.34 (1997), pp. 10581–10594.

[6] Johan Banér et al. "Signal amplification of padlock probes by rolling circle replication". In: *Nucleic Acids Research* 26.22 (1998), pp. 5073–5078. ISSN: 0305-1048.

[7] Anton Bankevich et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". eng. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19.5 (May 2012), pp. 455–477. ISSN: 1557-8666.

[8] Paul C. Blainey. "The future is now: single-cell genomics of bacteria and archaea". In: *FEMS Microbiology Reviews* 37.3 (May 2013), pp. 407–427. ISSN: 0168-6445.

[9] Luis Blanco, Antonio Bernad, and Margarita Salas. *Phi29 DNA Polymerase*. US Patent 5,198,543. 1993.

[10] Scott D. Boyd et al. "Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements". en. In: *The Journal of Immunology* 184.12 (June 2010), pp. 6986–6992. ISSN: 0022-1767, 1550-6606.

[11] Robert K Bradley et al. "Fast statistical alignment". In: *PLoS computational biology* 5.5 (2009), e1000392.

[12] J. J. Calis and Rosenberg BR. "Characterizing immune repertoires by high throughput sequencing: strategies and applications". In: *Trends Immunol* 35.12 (2014), pp. 581–590.

[13] P. Campbell et al. "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing". In: *Nat Genet* 40 (2008), pp. 722–729.

[14] John P. Cannon et al. "The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms". en. In: *Immunological Reviews* 200.1 (Aug. 2004), pp. 12–22. ISSN: 1600-065X.

[15] D. Chiang et al. "High-resolution mapping of copy-number alterations with massively parallel sequencing". In: *Nat Methods* 6 (2009), pp. 99–103.

[16] N.-O. Chimge et al. "Determination of gene organization in the human IGHV region on single chromosomes". eng. In: *Genes and Immunity* 6.3 (May 2005), pp. 186–193. ISSN: 1466-4879.

[17] C. S. Cho et al. "Genotyping by PCR-ELISA of a complex polymorphic region that contains one to four copies of six highly homologous human VH3 genes". eng. In: *Proceedings of the Association of American Physicians* 109.6 (Nov. 1997), pp. 558–564. ISSN: 1081-650X.

[18] Erik L. Clarke et al. "swga: a primer design toolkit for selective whole genome amplification". In: *Bioinformatics* 33.14 (2017), pp. 2071–2077. ISSN: 1367-4803.

[19] Peter J. A. Cock et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". en. In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. ISSN: 1367-4803.

[20] Donald F. Conrad et al. "A worldwide survey of haplotype variation and linkage disequilibrium in the human genome". eng. In: *Nature Genetics* 38.11 (Nov. 2006), pp. 1251–1260. ISSN: 1061-4036.

[21] G. P. Cook et al. "A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14q". In: *Nat Genetics* 7.2 (1994), pp. 162–168.

[22] Martin M. Corcoran et al. "Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity". en. In: *Nature Communications* 7 (Dec. 2016), p. 13642. ISSN: 2041-1723.

[23] Annie N. Cowell et al. "Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed Clinical Samples". In: *mBio* 8.1 (2017), e02257–16.

[24] Sabyasachi Das et al. "Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates". en. In: *Immunogenetics* 60.1 (Jan. 2008), pp. 47–55. ISSN: 0093-7711, 1432-1211.

[25] Frank B. Dean et al. "Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification". In: 11.6 (2001), pp. 1095–1099.

[26] C. W. Dieffenbach, T. M. Lowe, and G. S. Dveksler. "General concepts for PCR primer design". In: *Genome Research* 3.3 (1993), S30–S37. ISSN: 1054-9805/93.

[27] N. A. Doria-Rose et al. "Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies". In: *Nature* 509.7498 (2014), pp. 55–62.

[28] A. J. Drummond et al. "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Mol Biol Evol* 29.8 (2012), pp. 1969–1973.

[29] Michael A Eberle et al. "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". In: *Genome research* 27.1 (2017), pp. 157–164.

[30] Peter Edge, Vineet Bafna, and Vikas Bansal. "HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies". In: *Genome research* 27.5 (2017), pp. 801–812.

[31] Jonathan A. Eisen. "Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes". In: *PLOS Biology* 5.3 (2007), e82.

[32] R. O. Emerson et al. "Robust Detection Of Minimal Residual Disease In Unselected Patients With B-Cell Precursor Acute Lymphoblastic Leukemia By High-Throughput Sequencing Of IGH". In: *Blood* 122.21 (2013), pp. 2550–2550.

[33] Martin F Flajnik and Masanori Kasahara. "Origin and evolution of the adaptive immune system: genetic events and selective pressures". In: *Nature reviews. Genetics* 11.1 (Jan. 2010), pp. 47–59. ISSN: 1471-0056.

[34] Daniel Gadala-Maria et al. "Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles". en. In: *Proceedings of the National Academy of Sciences* 112.8 (Feb. 2015), E862–E870. ISSN: 0027-8424, 1091-6490.

[35] George Georgiou et al. "The promise and challenge of high-throughput sequencing of the antibody repertoire". en. In: *Nature Biotechnology* 32.2 (Feb. 2014), pp. 158–168. ISSN: 1087-0156.

[36] Moriah Gidoni et al. "Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping". En. In: *Nature Communications* 10.1 (Feb. 2019), p. 628. ISSN: 2041-1723.

[37] Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. "IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes". In: *Nucleic Acids Research* 33.suppl_1 (Jan. 2005), pp. D256–D261. ISSN: 0305-1048.

[38] Ann M. Guggisberg et al. "Whole-Genome Sequencing to Evaluate the Resistance Landscape Following Antimalarial Treatment Failure With Fosmidomycin-Clindamycin". In: *The Journal of Infectious Diseases* 214.7 (2016), pp. 1085–1091. ISSN: 0022-1899.

[39] W. Huang et al. "ART: a next-generation sequencing read simulator". In: *Bioinformatics* 28.4 (2012), pp. 593–594.

[40] Jin Inoue, Tsutomu Mikawa, and Yasushi Shigemori. "Improvements of rolling circle amplification (RCA) efficiency and accuracy using Thermus thermophilus SSB mutant protein". In: *Nucleic Acids Research* 34.9 (2006), e69–e69. ISSN: 0305-1048.

[41] K. J. Jackson et al. "Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements". In: *Cell Host Microbe* 16.1 (2014), pp. 105–114.

[42] Mattias Jakobsson et al. "Genotype, haplotype and copy-number variation in worldwide human populations". eng. In: *Nature* 451.7181 (Feb. 2008), pp. 998–1003. ISSN: 1476-4687.

[43] N. Jiang et al. "High Throughput Sequencing of the Human Antibody Repertoire in Response to Influenza Vaccination". In: *J Immunol* 188 (2012), pp. 58–14.

[44] N. Jiang et al. "Lineage structure of the human antibody repertoire in response to influenza vaccination". In: *Sci Transl Med* 5 (2013), p. 171.

[45] Shuai Jin et al. "Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice". In: *Science* 364.6437 (2019), pp. 292–295. ISSN: 0036-8075.

[46] J. M. Kidd et al. "A human genome structural variation sequencing resource reveals insights into mutational mechanisms". In: *Cell* 143.5 (2010), pp. 837–847.

[47] J. M. Kidd et al. "Mapping and sequencing of structural variation from eight human genomes". In: *Nature* 453.7191 (2008), pp. 56–64.

[48] M. J. Kidd et al. "The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements". In: *J Immunol* 188.3 (2012), pp. 1333–1340.

[49] D. C. Koboldt et al. "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome Res* 22 (2012), pp. 568–576.

[50] Victor Kunin et al. "A Bioinformatician's Guide to Metagenomics". In: *Microbiology and Molecular Biology Reviews* 72.4 (2008), p. 557.

[51] B. Langmead and Salzberg SL. "Fast gapped-read alignment with Bowtie 2". In: *Nat Methods* 9.4 (2012), pp. 357–359.

[52] K. Larimore et al. "Shaping of human germline IgH repertoires revealed by deep sequencing". In: *J Immunol* 189.6 (2012), pp. 3221–3230.

[53] Roger S. Lasken and Jeffrey S. McLean. "Recent advances in genomic DNA sequencing of microbial species from single cells". In: *Nature Reviews Genetics* 15 (2014), pp. 577–584.

[54] Marie-Paule Lefranc et al. "IMGT, the international ImMunoGeneTics database". In: *Nucleic acids research* 27.1 (1999), pp. 209–212.

[55] Aaron R. Leichty and Dustin Brisson. "Selective Whole Genome Amplification for Resequencing Target Microbial Species from Complex Natural Samples". In: *Genetics* 198.2 (2014), pp. 473–481.

[56] Samuel Levy et al. "The diploid genome sequence of an individual human". In: *PLoS biology* 5.10 (2007), e254.

[57] H. Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[58] Honghua Li et al. "Genetic diversity of the human immunoglobulin heavy chain VH region". en. In: *Immunological Reviews* 190.1 (2002), pp. 53–68. ISSN: 1600-065X.

[59] Jun Z. Li et al. "Worldwide human relationships inferred from genome-wide patterns of variation". eng. In: *Science (New York, N.Y.)* 319.5866 (Feb. 2008), pp. 1100–1104. ISSN: 1095-9203.

[60] H. X. Liao et al. "Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus". In: *Nature* 496.7446 (2013), pp. 469–476.

[61] L. Liu and Lucas AH. igh V3-23*01 and. "and its allele V3-23*03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of Haemophilus influenzae type b". In: *Immunogenetics* 55.5 (2003), pp. 336–338.

[62] Li et al. *Utilization of Ig heavy chain variable, diversity, and joining gene segments in children with B-lineage acute lymphoblastic leukemia: implications for the mechanisms of VDJ recombination and for pathogenesis — Blood Journal*. June 2004.

[63] Dorothy E. Loy et al. "Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites". In: *Proceedings of the National Academy of Sciences* 115.36 (2018), E8450–E8459.

[64] Shishi Luo, Jane A. Yu, and Yun S. Song. "Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads". In: *PLOS Computational Biology* 12.9 (Sept. 2016), e1005117. ISSN: 1553-7358.

[65] Shishi Luo et al. "Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans". In: *Life Science Alliance* 2.2 (2019).

[66] Mackelprang, Rachel, Carlson, Christopher S., and Subrahmanyan, Lakshman. "Sequence variation in the human T-cell receptor loci". In: *Immunological Reviews* (Dec. 2002). ISSN: 0105-2896.

[67] Swapan Mallick et al. "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". en. In: *Nature* 538.7624 (Oct. 2016), pp. 201–206. ISSN: 0028-0836.

[68] Lira Mamanova et al. "Target-enrichment strategies for next-generation sequencing". In: *Nature Methods* 7 (2010), pp. 111–118.

[69] Guillaume Marçais and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6 (2011), pp. 764–770.

[70] Elaine R. Mardis. "Next-Generation DNA Sequencing Methods". In: *Annual Review of Genomics and Human Genetics* 9.1 (2008), pp. 387–402. ISSN: 1527-8204.

[71] Fumihiko Matsuda et al. "The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus". en. In: *Journal of Experimental Medicine* 188.11 (Dec. 1998), pp. 2151–2162. ISSN: 0022-1007, 1540-9538.

[72] Aaron McKenna et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome research* 20.9 (2010), pp. 1297–1303.

[73] John J Miles, Daniel C Douek, and David A Price. "Bias in the alpha, beta T-cell repertoire: implications for disease pathogenesis and vaccination". In: *Immunology and cell biology* 89.3 (2011), pp. 375–387.

[74] Eric C. B. Milner et al. "Polymorphism and Utilization of Human VH Genesa". en. In: *Annals of the New York Academy of Sciences* 764.1 (Sept. 1995), pp. 50–61. ISSN: 1749-6632.

[75] Lefranc MP. "IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF". In: *Cold Spring Harbor Protocols* 2011 (2011), p. 6.

[76] Kenneth Murphy and Casey Weaver. *Janeway's Immunobiology, 9th edition*. en. Garland Science, Mar. 2016. ISBN: 978-1-315-53324-7.

[77] Murtagh, F. *hclust*. `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html`. Online; accessed 10 October 2019.

[78] A. Arul Murugan et al. *Statistical inference of the generation probability of T-cell receptors from sequence repertoires.* en. 2012.

[79] Masatoshi Nei and Alejandro P. Rooney. "Concerted and Birth-and-Death Evolution of Multigene Families". In: *Annual review of genetics* 39 (2005), pp. 121–152. ISSN: 0066-4197.

[80] Y. Niimura and M. Nei. "Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates". In: *J Hum Genet* 51.6 (2006), pp. 505–517.

[81] Patrick Nosil and Alex Buerkle. "Population Genomics". In: *Nature Education Knowledge* 3.10 (2010), p. 8.

[82] David Olivieri et al. "An automated algorithm for extracting functional immunologic V-genes from genomes in jawed vertebrates". eng. In: *Immunogenetics* 65.9 (Sept. 2013), pp. 691–702. ISSN: 1432-1211.

[83] T. Ota and M. Nei. "Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family." In: *Molecular Biology and Evolution* 11.3 (May 1994), pp. 469–482. ISSN: 0737-4038.

[84] Samuel O. Oyola et al. "Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification". In: *Malaria Journal* 15.1 (2016), p. 597. ISSN: 1475-2875.

[85] A. Pain et al. "The genome of the simian and human malaria parasite Plasmodium knowlesi". In: *Nature* 455 (2008), pp. 799–803.

[86] Robert Pinard et al. "Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing". In: *BMC Genomics* 7.1 (2006), p. 216. ISSN: 1471-2164.

[87] Sreemanta Pramanik and Honghua Li. "Direct Detection of Insertion/Deletion Polymorphisms in an Autosomal Region by Analyzing High-Density Markers in Individual Spermatozoa". English. In: *The American Journal of Human Genetics* 71.6 (Dec. 2002), pp. 1342–1352. ISSN: 0002-9297, 1537-6605.

[88] Sreemanta Pramanik et al. "Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region". eng. In: *BMC genomics* 12 (Jan. 2011), p. 78. ISSN: 1471-2164.

[89] *PREMIER Biosoft PCR Primer Design Guidelines*. `http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html`.

[90] Pengfei Qin et al. "Quantitating and Dating Recent Gene Flow between European and East Asian Populations". En. In: *Scientific Reports* 5 (Apr. 2015), p. 9500. ISSN: 2045-2322.

[91] Sohini Ramachandran et al. "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.44 (Nov. 2005), pp. 15942–15947. ISSN: 0027-8424.

[92] Mark F. Richardson et al. "Population genomics of the Wolbachia endosymbiont in Drosophila melanogaster". In: *PLOS Genetics* 8.12 (2012), e1003129.

[93] Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. "DSK: k-mer counting with very low memory usage". In: *Bioinformatics* 29.5 (2013), pp. 652–653.

[94] H. Robins et al. "Ultra-sensitive detection of rare T cell clones". In: *J Immunol Methods* 375.1 (2012), pp. 14–19.

[95] Harlan Robins. "Immunosequencing: applications of immune repertoire deep sequencing". In: *Current Opinion in Immunology* 25.5 (Oct. 2013), pp. 646–652. ISSN: 0952-7915.

[96] Lee Rowen, Ben F. Koop, and Leroy Hood. "The complete 685-kilobase DNA sequence of the human beta T cell receptor locus". English. In: *Science; Washington* 272.5269 (June 1996), p. 1755. ISSN: 00368075.

[97] Gavin G. Rutledge and Cristina V. Ariani. "Finding the needle in the haystack". In: *Nature Reviews Microbiology* 15 (2017), p. 136.

[98] John SantaLucia Jr. "Physical Principles and Visual-OMP Software for Optimal PCR Design". In: *PCR Primer Design*. Ed. by Anton Yuryev. Totowa, NJ: Humana Press, 2007, pp. 3–33. ISBN: 978-1-59745-528-2.

[99] John SantaLucia Jr and Donald Hicks. "The Thermodynamics of DNA Structural Motifs". In: *Annual Review of Biophysics and Biomolecular Structure* 33.1 (2004), pp. 415–440.

[100] E. H. Sasso, T. Johnson, and T. J. Kipps. "Expression of the immunoglobulin VH gene 51p1 is proportional to its germline gene copy number." en. In: *The Journal of Clinical Investigation* 97.9 (May 1996), pp. 2074–2080. ISSN: 0021-9738.

[101] E. H. Sasso et al. "A fetally expressed immunoglobulin VH1 gene belongs to a complex set of alleles." en. In: *The Journal of Clinical Investigation* 91.6 (June 1993), pp. 2358–2367. ISSN: 0021-9738.

[102] E. Sasso, J. Buckner, and L. Suzuki. "Ethnic differences in VH gene polymorphism". In: *Ann N Y Acad Sci* 764.1 (1995), pp. 72–73.

[103] Cathrine Scheepers et al. "Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire". eng. In: *Journal of Immunology (Baltimore, Md.: 1950)* 194.9 (May 2015), pp. 4371–4378. ISSN: 1550-6606.

[104] Christel Schmeisser, Helen Steele, and Wolfgang R. Streit. "Metagenomics, biotechnology with non-culturable microbes". In: *Applied Microbiology and Biotechnology* 75.5 (2007), pp. 955–962. ISSN: 1432-0614.

[105] E. Seboun et al. "Unusual organization of the human T-cell receptor beta-chain gene complex is linked to recombination hotspots". en. In: *Proceedings of the National Academy of Sciences* 90.11 (June 1993), pp. 5026–5029. ISSN: 0027-8424, 1091-6490.

[106] Helena M.B. Seth-Smith et al. "Whole-genome sequences of Chlamydia trachomatis directly from clinical samples without culture". In: *Genome Research* 23.5 (2013), pp. 855–866.

[107] Scott T. Small et al. "Human Migration and the Spread of the Nematode Parasite Wuchereria bancrofti". In: *bioRxiv* (2018), p. 421248.

[108] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences". In: *Journal of Molecular Biology* 147.1 (Mar. 1981), pp. 195–197. ISSN: 0022-2836.

[109] Edward J. Steele and Sally S. Lloyd. "Soma-to-germline feedback is implied by the extreme polymorphism at IGHV relative to MHC". en. In: *BioEssays* 37.5 (May 2015), pp. 557–569. ISSN: 1521-1878.

[110] Subrahmanyan, Lakshman et al. "Sequence Variation and Linkage Disequilibrium in the Human T-Cell Receptor Beta (TCRB) Locus". In: *American Jounal of Human Genetics* 69 (2001), pp. 381–395.

[111] Peter H. Sudmant et al. "Global diversity, population stratification, and selection of human copy number variation". In: *Science (New York, N.Y.)* 349.6253 (Sept. 2015), aab3761. ISSN: 0036-8075.

[112] Sesh A. Sundararaman et al. "Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria". In: *Nature Communications* 7.7 (2016), p. 11078.

[113] K. Tamura and M. Nei. "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees". In: *Mol Biol Evol* 10.3 (1993), pp. 512–526.

[114] Sarah A. Tishkoff and Kenneth K. Kidd. "Implications of biogeography of human populations for 'race' and medicine". eng. In: *Nature Genetics* 36.11 Suppl (Nov. 2004), S21–27. ISSN: 1061-4036.

[115] Sarah A. Tishkoff and Brian C. Verrelli. "Patterns of human genetic diversity: implications for human evolutionary history and disease". eng. In: *Annual Review of Genomics and Human Genetics* 4 (2003), pp. 293–340. ISSN: 1527-8204.

[116] Sarah A. Tishkoff and Scott M. Williams. "Genetic analysis of African populations: human evolution and complex disease". eng. In: *Nature Reviews. Genetics* 3.8 (Aug. 2002), pp. 611–621. ISSN: 1471-0056.

[117] Eray Tuzun et al. "Fine-scale structural variation of the human genome". en. In: *Nature Genetics* 37.7 (July 2005), pp. 727–732. ISSN: 1546-1718.

[118] Sijia Wang et al. "Genetic Variation and Population Structure in Native Americans". In: *PLOS Genetics* 3.11 (Nov. 2007), e185. ISSN: 1553-7404.

[119] Yan Wang et al. "Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants". In: *Immunogenetics* 63.5 (2011), pp. 259–265.

[120] C. Watson and F. Breden. "The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease". In: *Genes and Immunity* 13.5 (2012), pp. 363–373.

[121] Corey T. Watson et al. "Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation". In: *The American Journal of Human Genetics* 92.4 (Apr. 2013), pp. 530–546. ISSN: 0002-9297.

[122] Robinson WH. "Sequencing the functional antibody repertoire – diagnostic and therapeutic discovery". In: *Nat Rev Rheumatol* 11.3 (2015), pp. 171–182.

[123] X. Wu et al. "Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing". In: *Science* 333.6049 (2011), pp. 1593–1602.

[124] Jian Ye et al. "IgBLAST: an immunoglobulin variable domain sequence analysis tool". In: *Nucleic acids research* 41.W1 (2013), W34–W40.

[125] S. Yoon et al. "and accurate detection of copy number variants using read depth of coverage". In: *Genome Res* 19 (2009), pp. 1586–1592.

[126] B. Zhang et al. "Discrimination of germline V genes at different sequencing lengths and mutational burdens: A new tool for identifying and evaluating the reliability of V gene assignment". In: *J Immunol Methods* 427 (2015), pp. 105–116.

[127] Xian-Yang Zhang, Vincent F. La Russa, and Jakob Reiser. "Transduction of Bone-Marrow-Derived Mesenchymal Stem Cells by Using Lentivirus Vectors Pseudotyped with Modified RD114 Envelope Glycoproteins". In: *Journal of Virology* 78.3 (2004), pp. 1219–1229.

[128] Mengyao Zhao et al. "SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications". In: *PLOS ONE* 8.12 (Dec. 2013), e82138. ISSN: 1932-6203.

[129] Zhongming Zhao et al. "Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations". eng. In: *Genetics* 174.1 (Sept. 2006), pp. 399–409. ISSN: 0016-6731.

[130] Erwei Zuo et al. "Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos". In: *Science* 364.6437 (2019), pp. 289–292. ISSN: 0036-8075.

Appendix A

Chapter 2 Supplementary Information

A.1 Supplementary Figures

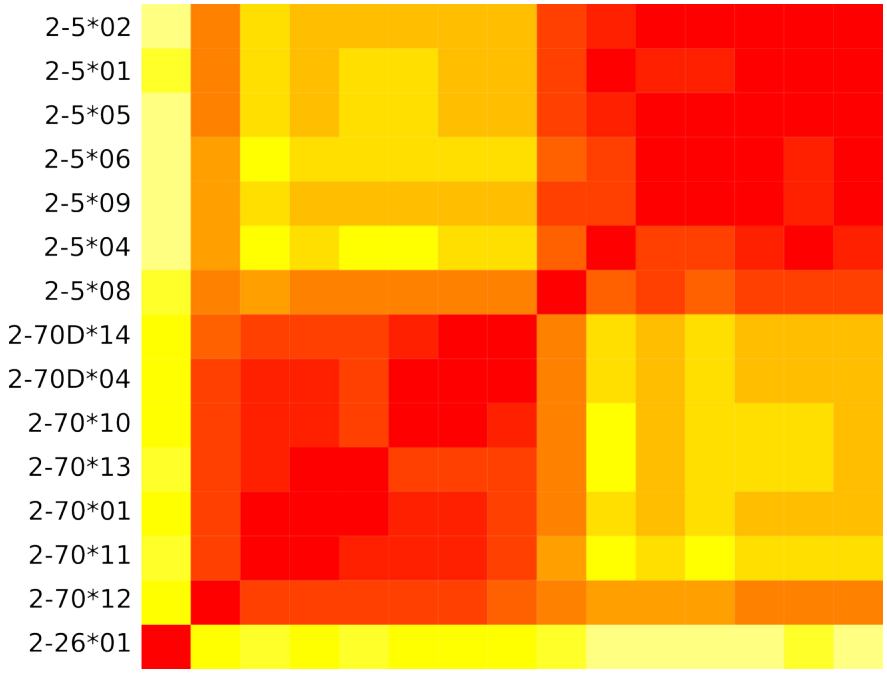


Figure A.1: Hierarchical clustering applied to Hamming distance between all family 2 alleles. Heatmap color scale is same as in the main text, with red=0% nucleotide differences, white=10% or more.

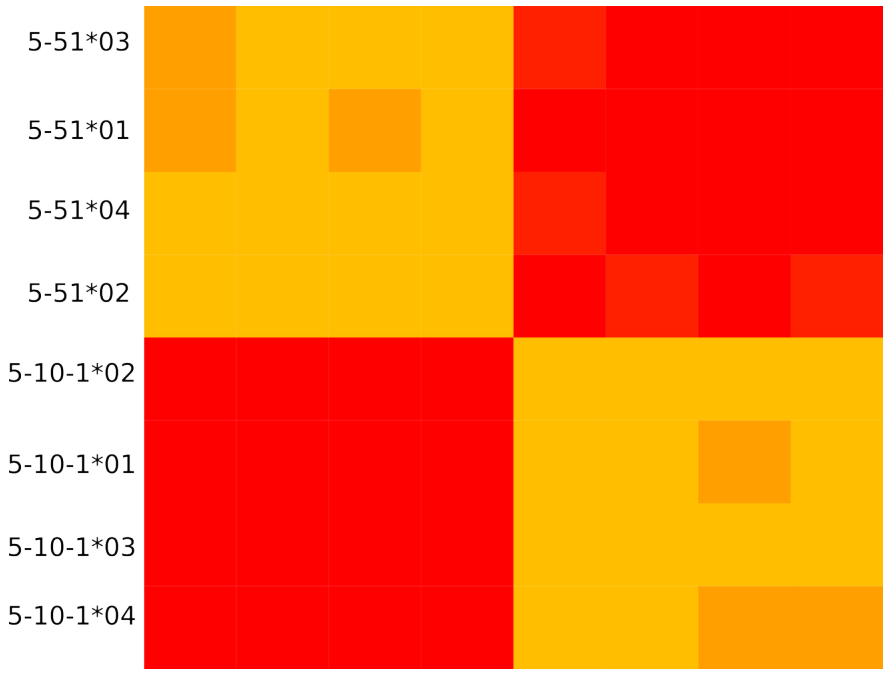


Figure A.2: Hierarchical clustering applied to Hamming distance between all family 5 alleles. Heatmap color scale is same as in the main text, with red=0% nucleotide differences, white=10% or more.

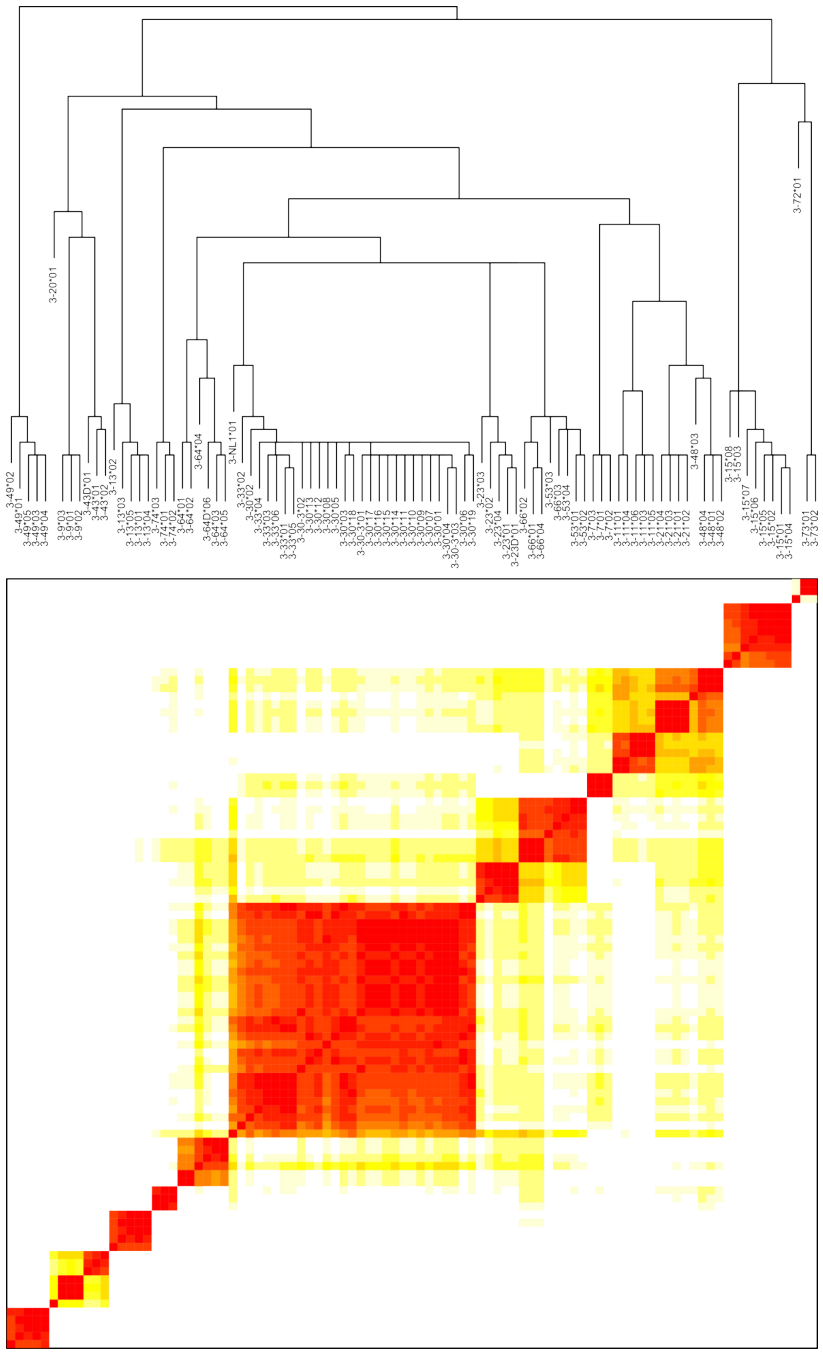


Figure A.3: Hierarchical clustering applied to Hamming distance between all family 3 alleles. Allele labels are in cladogram above matrix. Heatmap color scale is same as in the main text, with red=0% nucleotide differences, white=10% or more.

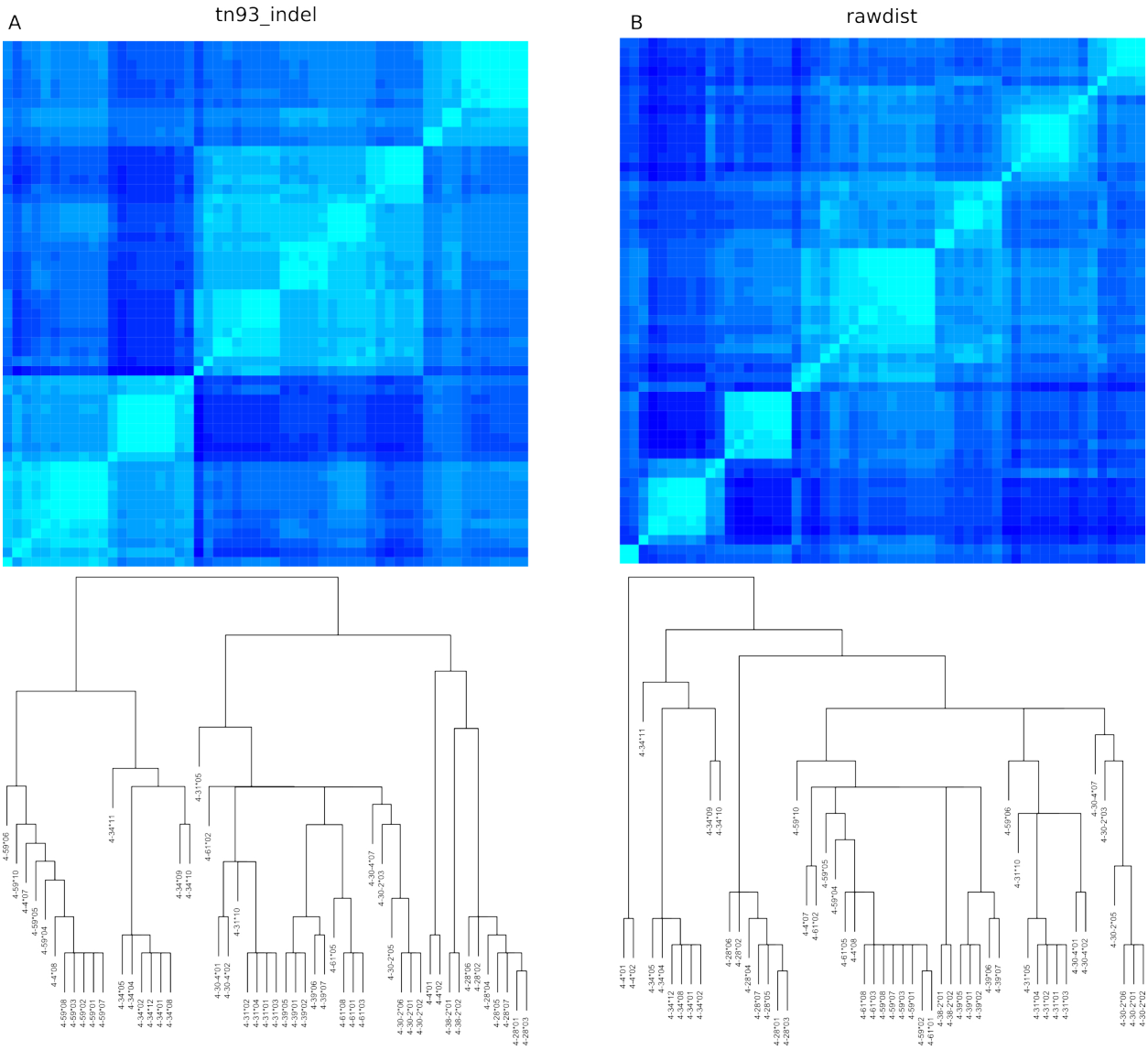


Figure A.4: Hierarchical clustering applied to Hamming distance between all family 4 alleles. (A) Simple average of 'TN93' evolutionary distance and indel distance. (B) Hamming distance. Allele labels are in cladogram below matrix. Because TN93 and indel distances cannot be interpreted in terms of nucleotide similarity, the distances in each matrix have been normalized by the maximum value in the matrix for comparison. Heatmap color scale ranges from cyan=0 to blue=1. The clustering that uses 'TN93' and 'indel' distances has clearer block diagonal structure and fewer conflicts with IMGT nomenclature. It was therefore used to define the operational segments for family 4 in Table 1.

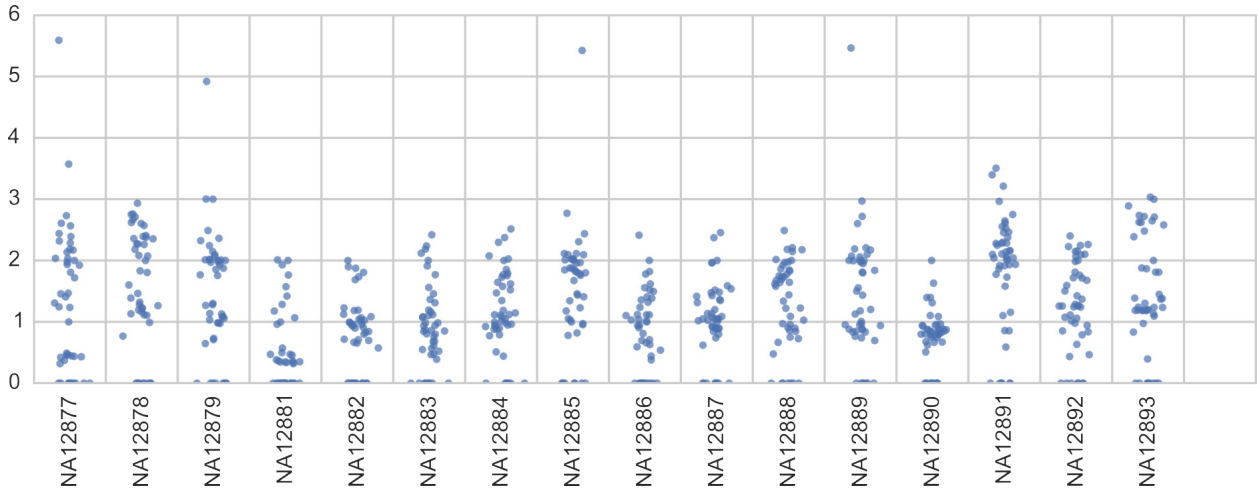


Figure A.5: Dotplots of coverage calls for each individual in the Platinum Genomes dataset. The data points are the same as in Fig. 4 but grouped by individuals. Y axis is normalized read coverage depth.

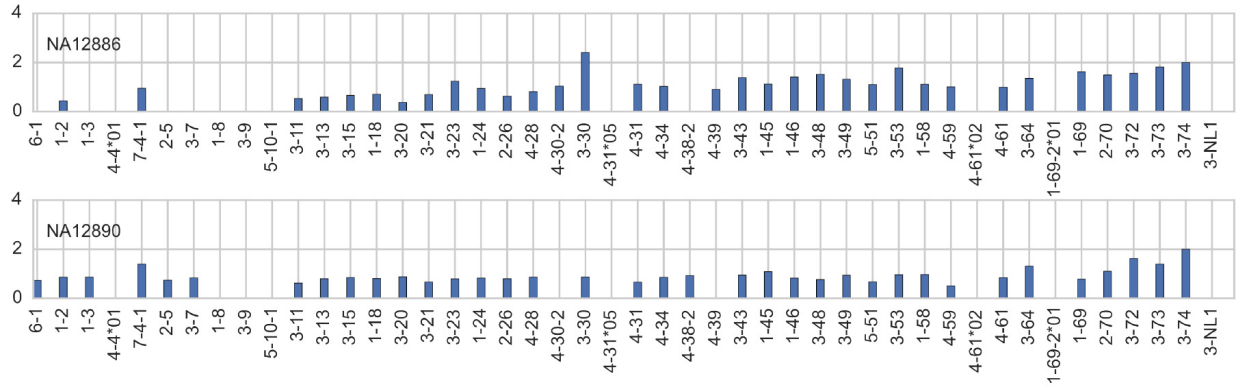


Figure A.6: Normalized coverage of subjects NA12886 and NA12890. The absence of *1-8/3-9* and *5-10-1/3-64D* variants does not appear to be due to VDJ recombination.

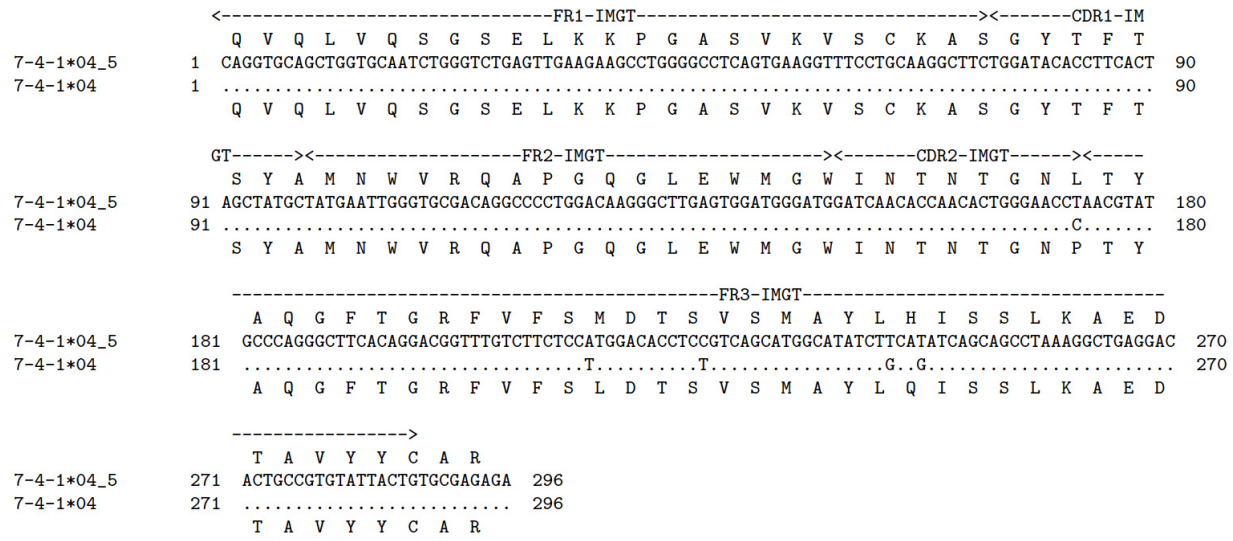


Figure A.7: Pairwise alignment of the putative *7-4-1* allele, *7-4-1∗04_5*, with its closest matching IMGT allele, *7-4-1∗04*. The allele *7-4-1∗04_5* was found in individuals NA12877, NA12878, NA12879, NA12883, NA12884, NA12886, NA12888, NA12891, and NA12893. Pairwise alignment was performed using the online IgBLAST tool [124].

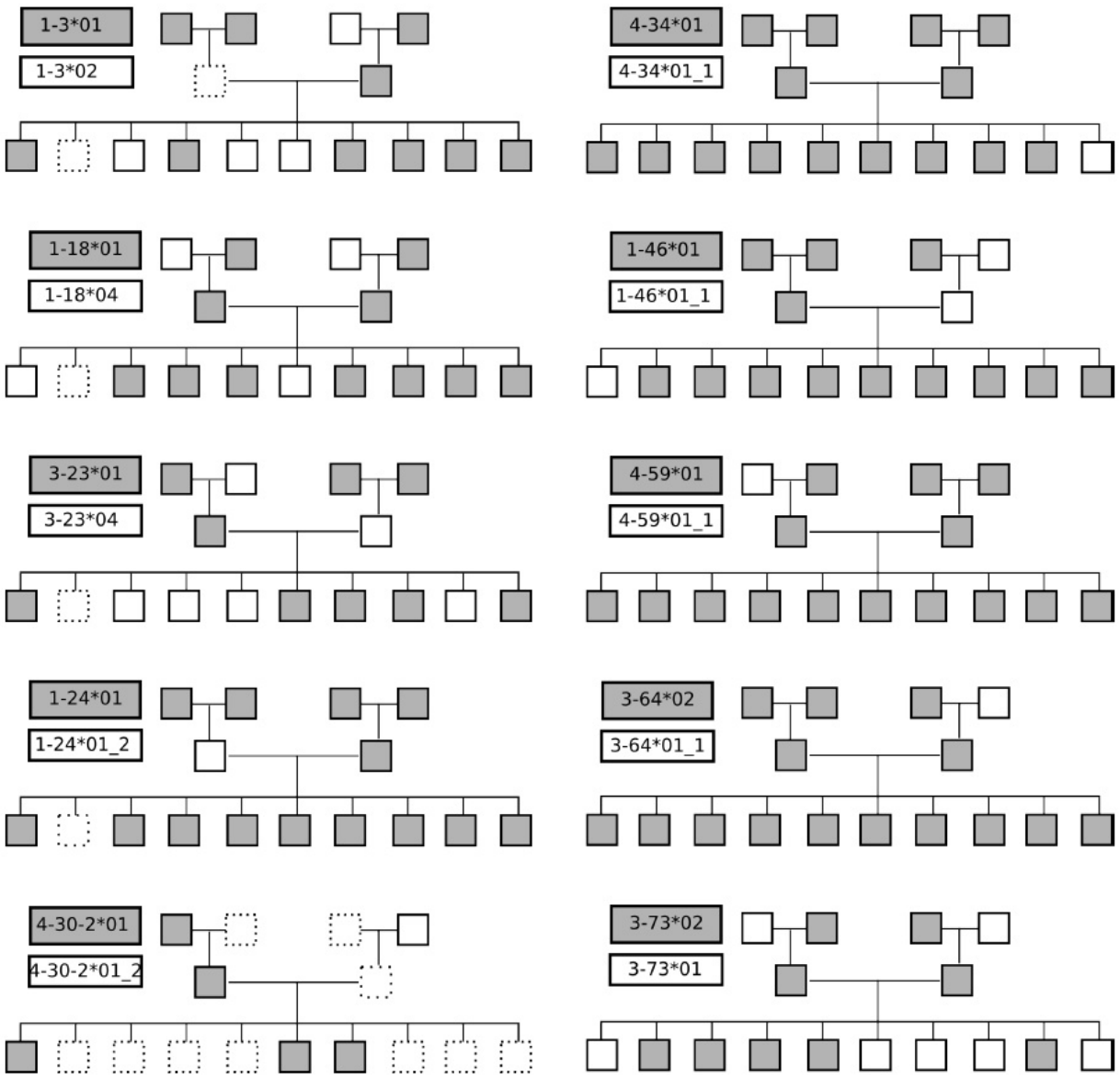


Figure A.8: Allele calls arranged according to family pedigree. Only segments for which there were two alleles in the family are shown (colored grey and white). Individuals who did not carry the segment are denoted by boxes with dashed outlines.

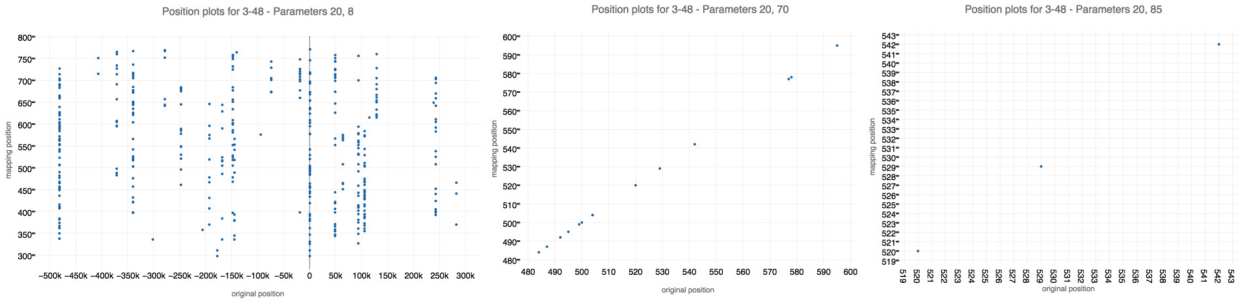


Figure A.9: Mapped start position versus original start position in segment *3-48* of each 250 bp read whose alignment exceeds the score threshold. Axis values are centered at position chr14:1,062,766,005. (A) With default Bowtie2 local alignment threshold of $20 + 8.0\ln(L)$, where L is the read length, reads originally from pseudogenes or similar functional segments are incorrectly mapped to *3-48*, as seen by multiple vertical strips of dots. (B) With the threshold increased to $20 + 70\ln(L)$, a single diagonal row of dots indicates that only reads from *3-48* are mapped to segment *3-48*. (C) When the threshold is increased to $20+85\ln(L)$ however, this is too restrictive and too few reads are mapped. Assessing analogous plots for the rest of the segments led to a threshold of $20 + 70\ln(L)$ being chosen. The README of the package provides more detail on how to modify the threshold. (Coordinates are for chromosome 14 on GRCh37).

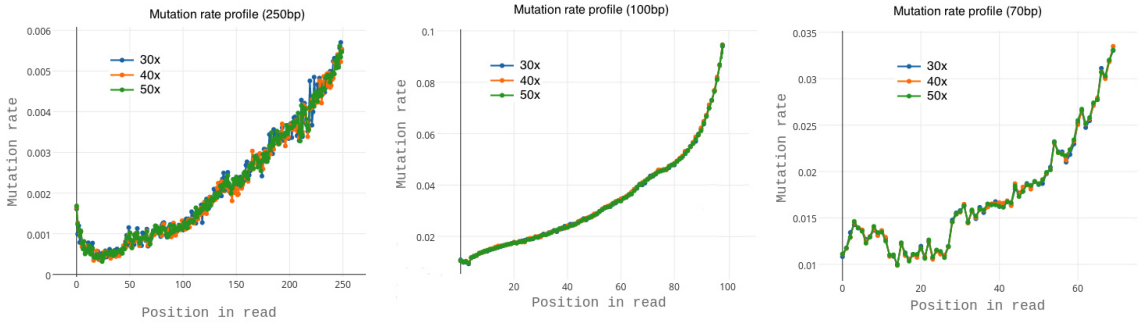


Figure A.10: Error profiles of simulated reads under default ART parameters. Plots are shown for 30x, 40x, and 50x coverages and are only displayed for GRCh37 (plots for GRCh38 are similar). Note the high error rates for 100 bp and 70 bp reads. This difference is attributed to the fact that ART automatically selects one of several built-in read quality profiles according to the read length provided. Mutation rates are computed by first calculating, for each position in the read, the number of mismatches between the position of the simulated nucleotide and the original nucleotide. The number of mismatches was then divided by the total number of reads.

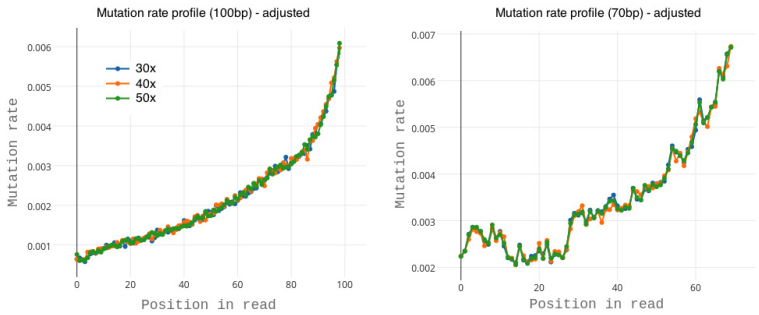


Figure A.11: Error profiles of simulated reads after parameter adjustment. To make the profiles for 70bp and 100bp comparable to that of 250bp, the parameter for quality score shifting (-qs and -qs2) was used: 12.896 for 100bp and 7.99 for 70bp.

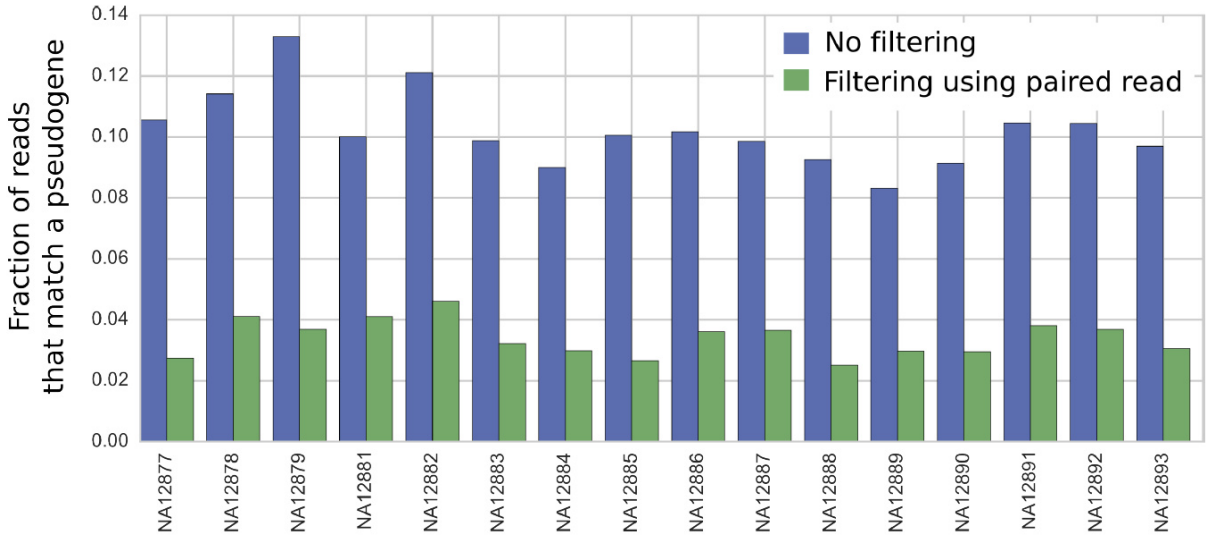


Figure A.12: Fraction of reads from an individual which matches a psuedogene before and after filtering using paired reads. Filtering using mate-pair information decreases reads from pseudogenes. For each individual, we calculated the fraction of reads that more closely match a pseudogene than an IMGT allele (blue). We then did the same calculation after filtering reads using the location of the paired read (green).

A.2 Supplementary Tables

Table A.1: Table of IMGT alleles to which simulated GRCh37 reads map ambiguously. Note that the in cases where reads map exactly to one allele, i.e. *3-72∗01*, *2-26∗01*, *1-24∗01*, and *3-20∗01*, these are the only full-length and functional alleles corresponding to a segment. Simulated reads are 100 bp long and have coverage depth of 30x.

Read from	Is mapped to
3-74∗01	3-74∗01, 3-74∗02, 3-74∗03
3-73∗02	3-73∗01, 3-73∗02
3-72∗01	3-72∗01
2-70∗13	2-70∗01, 2-70∗10, 2-70∗11, 2-70∗12, 2-70∗13
1-69∗06	1-69∗01, 1-69∗02, 1-69∗04, 1-69∗05, 1-69∗06, 1-69∗08, 1-69∗09, 1-69∗12, 1-69∗14
3-66∗03	3-30∗14, 3-53∗01, 3-53∗04, 3-66∗02, 3-66∗03
3-64∗02	3-64∗01, 3-64∗02
4-61∗08	4-59∗01, 4-59∗02, 4-59∗03, 4-59∗07, 4-61∗01, 4-61∗03, 4-61∗08
4-59∗01	4-4∗08, 4-59∗01, 4-59∗02, 4-59∗03, 4-59∗04, 4-59∗07, 4-59∗08, 4-61∗01, 4-61∗03, 4-61∗05, 4-61∗08
1-58∗02	1-58∗01, 1-58∗02
3-53∗01	3-53∗01, 3-53∗02, 3-53∗03, 3-53∗04, 3-66∗02, 3-66∗03
5-51∗01	5-51∗01, 5-51∗02, 5-51∗03, 5-51∗04
3-49∗03	3-49∗01, 3-49∗02, 3-49∗03, 3-49∗04, 3-49∗05
3-48∗02	3-13∗01, 3-13∗04, 3-23∗04, 3-48∗01, 3-48∗02, 3-48∗04
1-46∗01	1-46∗01, 1-46∗02, 1-46∗03, 3-11∗04, 3-11∗06, 3-48∗03, 3-48∗04, 3-66∗01, 3-7∗01, 3-7∗02, 3-7∗03
1-45∗02	1-45∗01, 1-45∗02
3-43∗01	3-43∗01, 3-43∗02, 3-43D∗01
4-39∗01	4-30-2∗03, 4-39∗01, 4-39∗02, 4-39∗05, 4-39∗06, 4-39∗07, 4-59∗05, 4-61∗05
4-34∗01	4-34∗01, 4-34∗02, 4-34∗04, 4-34∗05, 4-34∗08, 4-34∗12
3-33∗01	3-30∗02, 3-30∗06, 3-30∗07, 3-30∗11, 3-30∗12, 3-30-3∗02, 3-33∗01, 3-33∗02, 3-33∗03, 3-33∗04, 3-33∗05, 3-33∗06
4-31∗02	4-30-4∗01, 4-31∗01, 4-31∗02, 4-31∗03, 4-31∗04, 4-31∗05, 4-59∗06
3-30∗03	3-30∗02, 3-30∗03, 3-30∗04, 3-30∗05, 3-30∗06, 3-30∗07, 3-30∗10, 3-30∗13, 3-30∗17, 3-30∗18, 3-30-3∗02, 3-30-3∗03, 3-33∗01, 3-33∗05
4-28∗01	4-28∗01, 4-28∗02, 4-28∗03, 4-28∗04, 4-28∗05, 4-28∗07
2-26∗01	2-26∗01
1-24∗01	1-24∗01
3-23∗01	3-23∗01, 3-23∗02, 3-23∗03, 3-23∗04, 3-23D∗01
3-21∗01	3-11∗06, 3-21∗01, 3-21∗02, 3-21∗03, 3-21∗04
3-20∗01	3-20∗01
1-18∗01	1-18∗01, 1-18∗03, 1-18∗04
3-15∗01	3-15∗01, 3-15∗02, 3-15∗04, 3-15∗05, 3-15∗06, 3-15∗07
3-13∗01	3-13∗01, 3-13∗03, 3-13∗04, 3-13∗05, 3-23∗04, 3-48∗02, 3-7∗01, 3-7∗02, 3-7∗03
3-11∗01	3-11∗01, 3-11∗03, 3-11∗04, 3-11∗05, 3-11∗06
3-9∗01	3-9∗01, 3-9∗02, 3-9∗03
1-8∗01	1-2∗01, 1-2∗02, 1-2∗04, 1-8∗01, 1-8∗02
3-7∗01	1-46∗01, 1-46∗02, 1-46∗03, 3-21∗01, 3-21∗02, 3-64∗01, 3-7∗01, 3-7∗02, 3-7∗03
2-5∗01	2-5∗01, 2-5∗02, 2-5∗04, 2-5∗05, 2-5∗08, 2-5∗09
4-4∗07	4-39∗07, 4-4∗07, 4-4∗08, 4-59∗03, 4-59∗04, 4-59∗10, 4-61∗02
1-3∗02	1-3∗01, 1-3∗02
1-2∗02	1-2∗01, 1-2∗02, 1-2∗03, 1-2∗04, 1-2∗05
6-1∗01	6-1∗01, 6-1∗02

Table A.2: GRCh37 and GRCh38 in terms of our gene clusters. When there is one allele listed for a gene cluster, that gene cluster is considered to be in single copy. If there are two alleles listed, the gene cluster has two copies.

Gene cluster	GRCh37	GRCh38
6-1	6-1∗01	6-1∗01
1-2	1-2∗02	1-2∗04
1-3	1-3∗02	1-3∗01
4-4∗01	-	4-4∗02
7-4-1		7-4-1∗01
2-5	2-5∗01	2-5∗02
3-7	3-7∗01	3-7∗03
1-8	1-8∗01	-
3-9	3-9∗01	-
5-10-1	-	5-10-1∗03
3-11	3-11∗01	3-11∗06
3-13	3-13∗01	3-13∗05
3-15	3-15∗01	3-15∗01
1-18	1-18∗01	1-18∗04
3-20	3-20∗01	3-20∗02
3-21	3-21∗01	3-21∗01
3-23	3-23∗01	3-23∗04
1-24	1-24∗01	1-24∗01
2-26	2-26∗01	2-26∗01
4-28	4-28∗01	4-28∗07
4-30-2	-	4-30-2∗01
3-30	3-33∗01, 3-30∗03	3-30∗18, 3-33∗01
4-31	4-31∗02	-
4-34	4-34∗01	4-34∗01
4-39	4-39∗01	4-39∗01
3-43	3-43∗01	3-43∗01
1-45	1-45∗02	1-45∗02
1-46	1-46∗01	1-46∗01
3-48	3-48∗02	3-48∗03
3-49	3-49∗03	3-49∗04
5-51	5-51∗01	5-51∗01
3-53	3-53∗01, 3-66∗03	3-53∗02, 3-66∗03
1-58	1-58∗02	1-58∗01
4-59	4-4∗07, 4-59∗01	4-59∗01
4-61	4-61∗08	4-61∗01
3-64	3-64∗02	3-64∗02, 3-64D∗06
1-69-2	-	1-69-2∗01
1-69	1-69∗06	1-69D∗01, 1-69∗06
2-70	2-70∗13	2-70D∗04, 2-70∗01
3-72	3-72∗01	3-72∗01
3-73	3-73∗02	3-73∗02
3-74	3-74∗01	3-74∗01

Appendix B

Chapter 3 Supplementary Information

B.1 Read Mapping

For each individual's whole genome sequencing FASTQ file, we retained reads from two separate procedures: (i) reads that map to the IGHV and TRBV loci on the GRCh37 reference, (ii) reads that map to a list of functional IGHV and TRBV (from the online IMGT database).

Procedure (i) used bwa mem (`https://github.com/lh3/bwa`) with default parameters:

```
bwa mem grch37.fa read1.fq read2.fq | gzip -3 > aln-pe.sam.gz
```

Procedure (ii) used minimap (`https://github.com/lh3/minimap`):

```
minimap -w1 -f1e-9 imgt_ighv.fa.gz read-se.fa.gz > out_ighv.mini
minimap -w1 -f1e-9 imgt_trbv.fa.gz read-se.fa.gz > out_trbv.mini
```

Although procedure (i) was sufficient in our previous study [64] for identifying all the IGHV genes in an individual, it was clear that for the Simons dataset, the mapped reads were biased to the GRCh37 reference (Figure B.19). This is possibly due to differences in mapping algorithms in the different sample sets. For this reason, procedure (ii) was needed instead to 'catch' the reads that are not in the reference genome but which map to known IGHV and TRBV gene segments. Such gene segments for the GRCh37 reference genome are *IGHV7-4-1*, *IGHV4-4*, *IGHV5-10-1*, *IGHV4-30-2*, *IGHV4-30-4*, *IGHV4-38-2*, *IGHV1-69-2*, *IGHV3-NL1*, *TBRV5-8*, *TRBV6-2*, *TRBV7-2*, *TRBV7-7*, *TRBV7-9*, *TRBV10-3*, *TRBV11-3*, *TRBV12-3*, *TRBV12-5*, *TRBV13*, *TRBV14*, *TRBV15*, *TRBV16*, and *TRBV18*.

Thus, unless otherwise stated, all our results are based on reads that were mapped using procedure (ii) *only*.

B.2 Read Filtering

We will call the set of reads obtained via procedure (ii), R_{IMGT}. R_{IMGT} includes reads from all parts of an individual's genome (functional, pseudo, and orphon genes) that share some sequence similarity with the list of functional IGHV alleles [37]. Our goal is to remove reads from pseudogenes and orphon genes and also resolve instances where a single read maps equally well to more than one functional gene. We devised gene-specific filtering rules to minimize erroneously mapped reads.

Operationally indistinguishable IGHV genes

First, we established that some IGHV genes are operationally indistinguishable from each other. Specifically, IGHV genes at distinct genomic locations have alleles that are highly similar (more than 95% nucleotide similarity). These indistinguishable sets are:

- {*IGHV3-23*, *IGHV3-23D*}
- {*IGHV3-30*, *IGHV3-30-3*, *IGHV3-30-5*, *IGHV3-33*}
- {*IGHV3-43*, *IGHV3-43D*}
- {*IGHV3-53*, *IGHV3-66*}
- {*IGHV3-64*, *IGVH3-64D*}
- {*IGHV1-69*, *IGHV1-69D*}
- {*IGHV2-70*, *IGHV2-70D*}

The basis for grouping these IGHV genes together was detailed in [64]. Note, however, that here we do not use the operational clusters that mix IMGT segment labels so as to minimize confusion with IMGT nomenclature. For the purposes of our study, we do not attempt to discriminate between IGHV genes in the above sets. Taking the full set of 54 IMGT functional gene segments and combining those in the above sets gives a set of 45 operationally distinguishable functional IGHV segments.

Discarding reads that map uniquely to pseudogenes and orphon genes. Our filtering begins by performing IgBLAST on all these reads against an expanded set of IGHV alleles that includes orphon genes and pseudogenes. Once we obtain the results of the IgBLAST procedure, we first discard all reads for which all the top hits are alleles of a single orphon gene or pseudogene.

For example, consider the following read:

```
>HS2000-1266_146:7:1206:16772:59318/1
```

```
GCTTGAGTGGATGGGATGGATCAACACTTACAATGGTAACACAAACTACCCACAGAAGCT
CCAGGGCAGAGTCACCATGACCAGAGACACATCCACGAGC
```

This read matches the orphon gene *IGHV1/OR15-2∗01* exactly and is therefore likely to have come from the orphon gene. However, it was originally included in R_{IMGT} because it is similar to positions 132-191 of functional allele*IGHV1-18∗01*, deviating by four nucleotides. Having established through IgBLAST that there is little ambiguity about where this read comes from, we discard it.

Functional IGHV genes to which 100 bp reads map uniquely. The set of reads we have left, call it R_{IMGT_fcn}, consist of reads that either uniquely map to a functional IGHV gene, or map equally well to regions of functional and pseudogenes/orphon genes. The former category is most straightforward to deal with. The 16 (out of 45) operationally distinguishable functional IGHV genes for which the reads in R_{IMGT_fcn} can be unambiguously mapped are:

IGHV6-1, *IGHV3-9*, *IGHV5-10-1*, *IGHV1-18*, *IGHV3-20*, *IGHV1-24*, *IGHV2-26*, {*IGHV3-43*, *IGHV3-43D*}, *IGHV1-45*, *IGHV3-49**, *IGHV5-51*, *IGHV1-58*, {*IGHV1-69*, *IGHV1-69D*}, *IGHV1-69-2** , *IGHV3-72*, *IGHV3-73*.

*These gene segments have inflated coverage in a subregion, see Supplementary Information Figure 1 of [65].

Functional IGHV genes to which 100 bp reads are not uniquely mapped. For reads that map equally well to more than one IGHV gene (functional or otherwise), we do not have enough information to assign reads to genes with 100% confidence. If we assign a read to all the functional genes that are tied for the top hit, some functional genes will have extra reads assigned to them. If we only assign the reads with a unique top hit, then we will lose information on functional genes that share subsequences of substantial length with other genes. Compounding matters is the fact that the presence/absence and copy number of IGHV genes can vary from individual to individual, so that an approach that works in one individual may have a different effect in another.

To determine the rules for filtering out reads that would have the least error on average, we used the read coverage profile for each IGHV segment, aggregated over all individuals in our sample. We compare the coverage profiles for a given segment under two read filtering rules:

1. 'All top hits': keep all reads which have that segment as a top hit (unique or tied), or
2. 'Unique top hits': keep only reads which have that segment as a unique top hit.

As an example of a well-behaved case, see Figure B.1 for a profile of *IGHV6-1*. Note that the per-base read coverage, averaged over all 109 individuals (from blood/saliva samples),

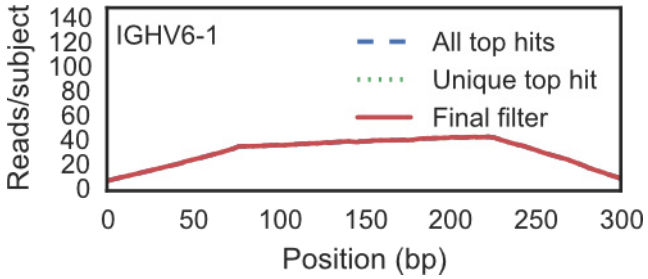


Figure B.1: Example of a well-behaved profile: *IGHV6-1*

matches very closely with what is expected theoretically. Specifically, the coverage decreases linearly towards the edges because reads that only partially cover the segment will have lower mapping scores and therefore were not in our original set R_{IMGT}. The per-base read coverage of around 40 is also consistent with the median genomic coverage of 42 across the full Simons sample (Supplementary Data Table 1 of [67]). This profile is evidence the set of reads that map to *IGHV6-1* does not contain reads from other similar IGHV genes, which is consistent with our earlier observation that *IGHV6-1* is a gene to which reads map uniquely. See Supplementary Information Figure 1 of [65] for the profiles of all the operationally distinguishable gene segments.

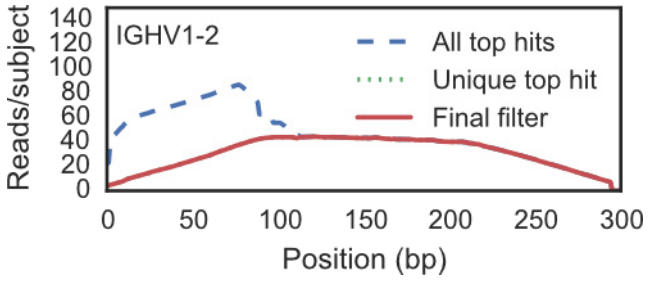


Figure B.2: Read coverage profile of *IGHV1-2*

We will describe a few representative examples of problematic cases here. See Figure B.2 for the profile for *IGHV1-2*. Some of the reads that map to *IGHV1-2* also map equally well (in lengths > 75 base pairs) with alleles of *IGHV1-8* and *IGHV1-OR15-1* (an orphon gene on chromosome 15). Indeed, the per-base read coverage of *IGHV1-2* average across our sample of 109 individuals shows that there is an overabundance of reads mapping to the first 100 base pairs of the gene. However, when we only keep the reads for which *IGHV1-2* is the unique top hit, the per-base read coverage more closely resembles what is expected theoretically. That this correction works suggests that reads mapping equally well to *IGHV1-2* and another gene in fact do not align with either very well, and therefore should be discarded as most likely not coming from *IGHV1-2*. Similar reasoning holds for *IGHV1-3*, *IGHV7-4-1*, *IGHV2-5*,

IGHV3-7, *IGHV1-8*, *IGHV3-11*, *IGHV3-13*, *IGHV3-21*, *IGHV3-23*, *IGHV4-30-2*, *IGHV4-30-4*, *IGHV3-30*, *IGHV4-34*, *IGHV1-46*, *IGHV3-48*, *IGHV3-53*, *IGHV3-64*, *IGHV2-70*, and *IGHV3-NL1*. It should be noted that we achieve varying levels of success. In particular for *IGHV3-15*, there was still quite elevated coverage after performing this step (see Table S1 of [65]). In other cases, we probably undercount the reads for some of these genes, giving us a conservative estimate of the segment copy number.

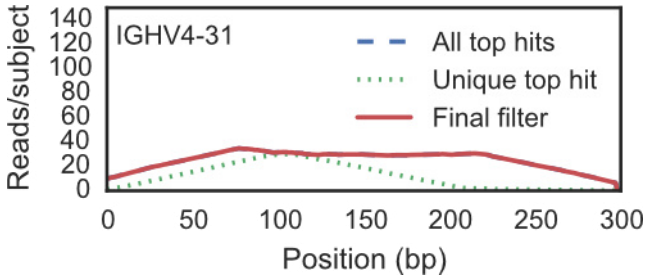


Figure B.3: Read coverage profile of *IGHV4-31*

See Figure B.3 for the profile for *IGHV4-31*. It was clear from the per-base coverage of *IGHV4-31* that if we discarded all the reads that mapped equally well to other genes, we would not obtain full sequence reconstruction of the gene even for the individuals it is present in. Moreover, counting all the reads that map to *IGHV4-31* did not suggest we were over counting. Thus, for *IGHV4-31*, we kept all the reads that had *IGHV4-31* as a top hit. This reasoning also held for *IGHV4-39* and *IGHV4-28*.

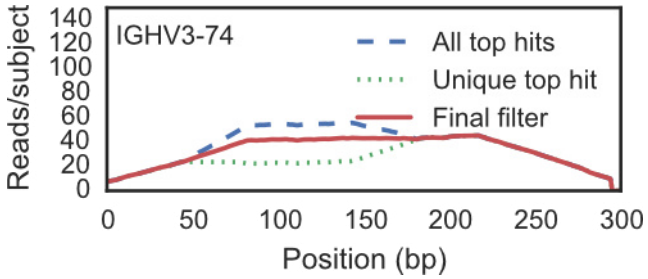


Figure B.4: Read coverage profile of *IGHV3-74*

See Figure B.4 for the profile for *IGHV3-74*. Among all the functional IGHV segments, *IGHV3-74* is unique in that it shares a long subsequence with an orphon gene that seems to be present in single copy on all haplotypes. Specifically, positions 46-182 (137 base pairs) of *IGHV3-74* (alleles ∗01 and ∗02, the two most common alleles) are identical to *IGHV1-OR/16-13∗01*. In this special case, we can toss a fair coin to determine whether to assign a read to *IGHV3-74* and be fairly confident we are counting accurately.

For *IGHV4-4*, *IGHV4-59*, *IGHV4-61*, and to a lesser extent *IGHV4-38-2* (see Figure B.5 for a representative profile), there are too many subregions that exactly match other IGHV genes. Counting only reads with a single top hit is a severe underestimate, and counting reads with ties for top hits is a severe overestimate. Thus, for lack of a better option, we sample the IGHV gene from the top hits at random, in proportion to the coverage of the top hits in the region outside the mapping region. Unlike the case with *IGHV3-74*, we know that the copy number of the other genes that share subsequences will differ between individuals. Hence, we know this approach is flawed. However, it gives our best estimate for the reads that come from these IGHV genes and our overall conclusions are not sensitive to the calls we make for these genes.

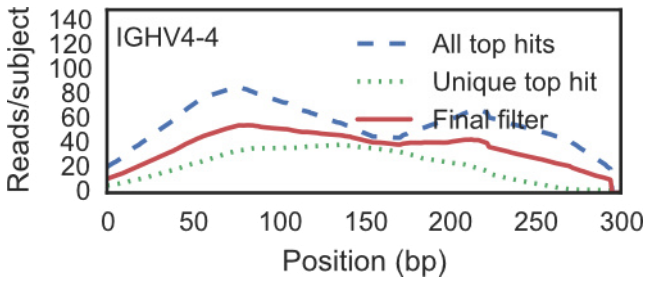


Figure B.5: Read coverage profile of *IGHV4-4*

After the above filtering steps have been performed, for each operationally distinguishable IGHV gene indexed by i, we have a set of reads, call it $R_{\text{filtered,i}}$.

Operationally indistinguishable TRBV genes

There were fewer issues in read filtering for the TRBV locus compared to the IGHV locus in general. As with the IGHV locus, there are some TRBV genes that are operationally indistinguishable from each other when using 100 bp reads. These are:

- {*TRBV4-2*, *TRBV4-3*}
- {*TRBV6-2*, *TRBV6-3*}
- {*TRBV12-3*, *TRBV12-4*}

For our purposes, we do not attempt to distinguish between segments within these sets. Taking the full set of 48 IMGT functional TRBV segments and combining those that are in the above sets gives 45 operationally distinguishable functional TRBV segments. Coincidentally, this is the same number of operationally distinguishable functional IGHV segments.

Discarding reads that map uniquely to pseudogenes and orphon genes. As with the IGHV reads, we begin by performing IgBLAST (for T cell receptors) on all the reads against an expanded set of TRBV alleles that includes orphon genes and pseudogenes. Once we obtain the results of the IgBLAST procedure, we first discard all reads for which the top hits are all alleles of a single orphon gene or pseudogene.

For example, consider the following read: `>HS2000-1266_146:7:1205:14671:84576/2`

```
GCTCCGGGCTTAGTGCTGTCGTCTCTCAACATCCGAGCAGGGTTATCTGTAAGAGTGGAA
CCTCTGTGAACATCGAGTGCCGTTCCCTGGACTTTCAGGC
```

This read matches the orphon gene *TRBV20/OR9-2∗01* exactly and is therefore likely to have come from the orphon gene. However, it was originally included in R_{IMGT} because it matches positions 1-89 of functional allele *TRBV20-1∗02*, deviating by two nucleotides. Having established through IgBLAST that there is little ambiguity about where this read comes from, we discard it.

Functional TRBV genes to which 100bp reads map uniquely. The set of reads we have left, call it R_{IMGT_fcn}, consist of reads that either uniquely map to a functional IGHV gene, or map equally well to regions of functional and pseudogenes/orphon genes. The former category is most straightforward to deal with. The 38 (out of 45) operationally distinguishable functional TRBV genes for which the reads in R_{IMGT_fcn} can be unambiguously mapped are:

TRBV2, *TRBV3-1*, *TRBV4-1*, {*TRBV4-2*, *TRBV4-3*}, *TRBV5-1*, *TRBV5-4*, *TRBV5-5*, *TRBV5-8*, {*TRBV6-2*, *TRBV6-3*}, *TRBV6-4*, *TRBV6-6*, *TRBV7-2*, *TRBV7-3*, *TRBV7-4*, *TRBV7-6*, *TRBV7-7*, *TRBV7-8*, *TRBV7-9*, *TRBV9*, *TRBV10-1*, *TRBV10-2*, *TRBV10-3*, *TRBV11-1*, *TRBV11-2*, *TRBV11-3*, {*TRBV12-3*, *TRBV12-4*}, *TRBV12-5*, *TRBV13*, *TRBV14*, *TRBV15*, *TRBV16*, *TRBV18*, *TRBV19*, *TRBV20-1*, *TRBV27*, *TRBV28*, *TRBV29-1*, *TRBV30*

Functional TRBV genes to which 100 bp reads are not uniquely mapped. As with the IGHV genes, we also have TRBV genes for which we cannot determine the correct reads with 100% confidence. This is a much smaller set of genes and again we compare the coverage profiles for a given TRBV gene under two read filtering rules:

1. 'All top hits': keep all reads which have that segment as a top hit (unique or tied), or
2. 'Unique top hits': keep only reads which have that segment as a unique top hit.

Supplementary Information Figure 2 of [65] contains the profiles of all the TRBV genes with summaries of how we cleaned up the reads mapping to them. We will describe two representative examples here. See Figure B.6 for a representative example: *TRBV5-5*. As with *IGHV3-74*, taking all top hits versus taking only unique top hits gave coverage profiles

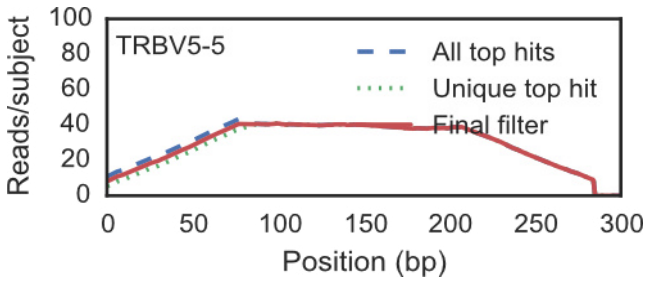


Figure B.6: Read coverage profile of *TRBV 5-5*

that were equally displaced above and below the theoretical expectation. Note that the displacement is much smaller than for *IGHV3-74*. Thus, for each read that has multiple top hits including *TRBV5-5*, we sample uniformly over the segments that are the top hits. This strategy was also used for *TRBV6-1*, *TRBV6-5*, *TRBV6-8*, *TRBV6-9*.

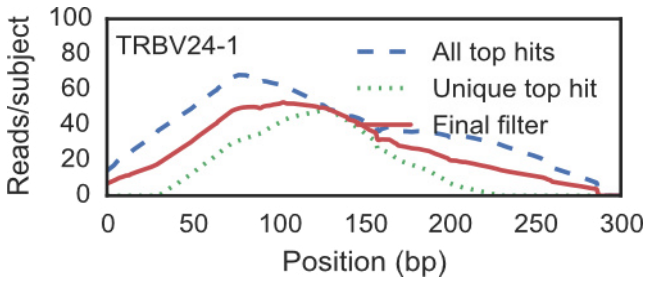


Figure B.7: Read coverage profile of *TRBV24-1*

See Figure B.7 of an example segment with a highly unusual coverage profile. For a lack of a better strategy, we also sampled uniformly over the segments that are top hits for reads with multiple top hits.

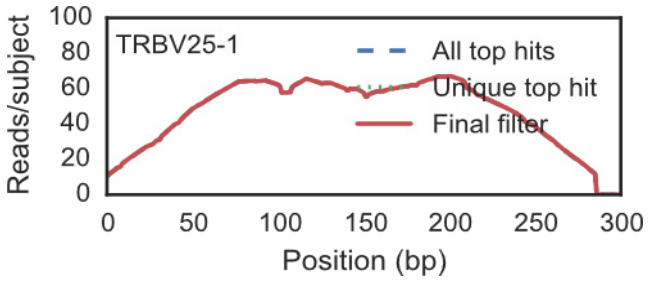


Figure B.8: Read coverage profile of *TRBV25-1*

See Figure B.8 for a last example profile. This segment also had a strange profile, but for a different reason. As with *IGHV3-15*, there seems elevated coverage, perhaps due to yet uncharacterized segments that share sequence similarity (Supplementary Information Figure 2 of [65]).

After the above filtering steps have been performed, for each operationally distinguishable TRBV gene indexed by i, we have a set of reads, call it $R_{\text{filtered,i}}$.

B.3 Copy number from kmer coverage

Suppose the kmer coverage of the reconstructed contig for a gene segment in an individual is d and the genome-wide coverage is g. Our point estimate for copy number would be:

$$c = d \times \frac{5}{4} \times \frac{4}{3} \times \frac{1}{g} \times 2$$

where 5/4 is the factor to convert 21-mer coverage of 100 bp to per-base coverage, 4/3 corrects for the fact that the trapezoidal coverage profiles (see Supplementary Information Figures 1 and 2 of [65]) gives per-base coverage depth that is $\frac{3}{4}$ of true uniform coverage depth, $1/g$ normalizes by the genome-wide average coverage depth, and we multiply by 2 so that a c of 2 corresponds to one copy per haplotype (two copies per individual).

However, we found that in practice, the distributions for copy number using this formula consistently led to underestimates of copy number across the segments. For example, the bulk of our point estimates for well-behaved and typically single copy per haplotype genes such as *IGHV6-1*, *IGHV2-5*, and *IGHV5-51*, were clustered just below the value two. The likely explanation is that the genome-wide coverage value we used is an overestimate of the true read coverage depth. We found that by multiplying the genome-wide coverage by 0.9, we obtained distributions for copy number that clustered more symmetrically around integer values, i.e., our point estimates were calculated as:

$$c = d \times \frac{5}{4} \times \frac{4}{3} \times \frac{1}{0.9g} \times 2$$

Finally, to obtain an integer from this point estimate, we simply round to the nearest whole number. The exception to this is when the point estimates look systematically biased across all the individuals, as is the case with *IGHV7-4-1*, *IGHV3-7*, *IGHV3-49*, and *IGHV1-69-2*. In these three cases, we calculate the mean shift of the points from integer values and move them all down by that shift before rounding to the nearest integer.

B.4 Hierarchical clustering

To call the copy number variants for common polymorphisms involving multiple genes, we use the point estimates for copy number. The reason we do not round the point estimates

to whole integers is that the rounding process may introduce additional error into our variant calls. In order to leverage knowledge of existing multi-gene CNVs to improve our copy number estimates, we performed hierarchical clustering (scipy.cluster.hierarchy) on the set of individuals, representing each individual as a vector comprised of the copy number estimate for each gene in the polymorphisms in Figure 3.2 and 3.3. For example, a scaled coverage value of 2.5 for *IGHV1-69* could be consistent with a copy number of 2 or 3, but if *IGHV2-70* has scaled coverage value of 2.1, it is more likely that *IGHV1-69* is copy number 2. Furthermore, given that *IGH1-69-2* had no reads mapping to it, we would be even more certain. Note that we did not use this method for any CNVs beyond those in Figure 3.2 and 3.3 involving two or more operationally distinguishable gene segments, because it would merely bin the copy number calls by intervals (e.g. Figure B.22B, Figure B.24B, Figure B.24C).

B.5 Determination of two-copy segments

To determine the set of 11 two-copy IGHV genes and 40 two-copy TRBV, we selected genes which are two copies in the vast majority of individuals in our sample and for which there is minimal read-mapping ambiguity. In other words, we selected genes for which the "Notes" column in Supplementary Information Figures 1 and 2 of [65] have "No subsequences shared with other known IGHV genes" and "Predominantly single copy per haplotype" .

Exceptions to this rule include *IGHV3-74*, *TRBV5-5*, *TRBV5-8*, *TRBV6-1*, *TRBV6-5*, *TRBV6-8*, and *TRBV6-9*. For these segments, the red solid line (the read profile for the filtered set of reads in Supplementary Information Figures 1 and 2 of [65]) lies roughly halfway between the green dotted line (the read profile resulting from taking the unique top hit) and the blue dashed line (the read profile resulting from taking all top hits). In such cases, the red line results from tossing a fair coin to determine which of the two genes to assign a given read. Because the resulting red read profile aligns with the expected trapezoidal shape, we believe that the alternative gene which shares the subsequence is present at the same copy number. Additionally, since the shared subsequence is identical in both genes, this should not lead to significant mis-mapping errors. *TRBV7-4* and *TRBV7-6* are also included in the set of two-copy genes since these two genes share a subsequence that is not 100% identical, and because the blue and green lines align, implying that the shared subsequence does not lead to mis-mapping. Consequently, we have included these operationally distinguishable gene segments in the set of two-copy segments, given that they are also predominantly single-copy per haplotype.

B.6 Alternative procedure for unphased variants from HapCUT2

When few or no reads covered two SNPs, HapCUT2 failed to phase the full segment. To account for this issue, we took all combinations of completely phased blocks as potential

allele sequences. As a toy example, consider a sequence of length three, each position having an unphased polymorphic site with A on one chromosome and G on the other. Taking combinations results in four pairs of haplotypes (AAA, GGG), (AGA, GAG), and (AAG, GGA), and (AGG, GAA). With the resulting pairs of phased sequences, we compute the probability of observing each candidate pair of allele sequences (a_1, a_2), which is calculated by taking the maximum of $P(a_1|\ a_2)P(a_2)$ and $P(a_2|\ a_1)P(a_1)$, where $P(a_x|\ a_y)$ is the fraction of individuals observed to have both allele sequences a_x and a_y out of all individuals observed to have allele a_y, and $P(a_x)$ is defined as the fraction of observations of a_x out of all individuals. If all candidate pairs of allele sequences were not observed in any other individual, then the following was selected as the individual's phased allele sequence pair: the pair of haplotypes that contains allele sequence a_x, which is the allele with the greatest frequency. If all candidate sequences a_x for an individual were not observed in the rest of the population, then no allele sequence was reported for that segment for that individual. This, however, does not affect our analysis of the presence/absence of gene segments.

B.7 Novel allele/SNV notation

Alleles were given IMGT names if their sequence exactly matched an allele in the IMGT database. Otherwise, the name of the closest allele was given with an appended suffix for each mutational difference from the closest IMGT allele. Each mutation is represented as {reference base pair} {alternative base pair} {position} {reference amino acid} {alternative amino acid}. For example, allele '*IGHV1-18∗01_ag168ND'* denotes an allele whose sequence is that of IMGT allele *IGHV1-18∗01*, but with a 'g' at position 168 rather than an 'a'. The two letters following the position are the amino acids corresponding to the reference base pair 'a' and to the mutation 'g', respectively (the reference amino acid is N and the alternate amino acid is D). If the sequence was equally close to more than one IMGT allele, the IMGT allele of the lowest numeric order was chosen. Alleles were called "novel" if it differed in at least one nucleotide from an existing IMGT allele. For SNVs, the notation is similar but mutations are represented simply as {alternative base pair} { position}.

In several tables, a 'P' and 'T' in parentheses indicates a stop codon and a truncation in one individual or more, respectively.

B.8 Method performance

The method utilized in this work is an extension of the method used in [64], which provides tests on simulated data and an application of the method to a sixteen-member pedigree of European descent. To summarize, we simulated reads using all combinations of read length (70, 100, 250bp), reference genomes (GRCh37 and GRCh38), and coverage depth (30x, 40x, 50x), and measured the recall, the fraction of operationally distinguishable gene segments

that are correctly called by the pipeline. All except 2 of the 18 combinations demonstrated 100% recall, and the remaining 2 simulations had a recall of 97%.

In this work, we have added steps towards haplotype phasing since the method from only constructed a single contig. In addition, we have now also performed simulations for both IGHV and TRBV. Simulating with the set of genes from 109 individuals for IGHV and 286 for TRBV that we empirically inferred from the SGDP dataset, we ran the reads through our pipeline to identify the accuracy rate. Specifically, out of all the alleles identified, we measured how many matched what was originally simulated. For IGHV this was 95.92% and for TRBV this was 98.43%. However, we emphasize again that other approaches may be more appropriate if the goal is to genotype a single individual at base-pair resolution, rather than a large set of individuals at a coarser resolution.

B.9 Analysis of IGHV and TRBV gene segments in 13 vertebrate species

Within-species analysis

For this analysis, we first measured the between-segment diversity of the nucleotide sequences annotated in the IMGT human gene table located at `http://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=human&group=IGHV`. In this study, we used only those functional alleles for which a position in the locus was recorded. Note that this resulted in a list of one allele sequence per gene segment. As a measure for diversity, we used pairwise global alignment with default BLASTN parameters (match=1, mismatch=–3, gap opening=–5, gap extension=–2). Alignment was done in python using the *pairwise2* module in the Biopython package [19]. Averaging over all possible pairs for IGHV gives a mean score of –44 and for TRBV this was –179.

We then expanded our analysis to thirteen vertebrate species, including human. For this analysis, we used amino acid sequences for IGHV segments and TRBV segments obtained from vgenereportoire.org [82]. For each species, the IGHV gene segments and TRBV gene segments were downloaded for one reference genome for that species. For reasons beyond our control, only the amino acid sequences and not the nucleotide sequences were available on the website. Species are: *homo sapiens* (human, `http://www.ncbi.nlm.nih.gov/nuccore/ABBA00000000.1`), *pan troglodytes* (chimpanzee, `http://www.ncbi.nlm.nih.gov/nuccore/AADA00000000.1`), *gorilla gorilla gorilla* (gorilla, `http://www.ncbi.nlm.nih.gov/nuccore/CABD000000000.3`), *pongo abelii* (orangutan, `http://www.ncbi.nlm.nih.gov/nuccore/ABGA00000000.1`), *macaca mulatta* (rhesus macaque, `http://www.ncbi.nlm.nih.gov/nuccore/AANU00000000.1`), *mus musculus* (mouse, `http://www.ncbi.nlm.nih.gov/nuccore/AAHY00000000.1`), *canis lupus familiaris* (dog, `http://www.ncbi.nlm.nih.gov/nuccore/AAEX00000000.3`), *oryctolagus cuniculus* (rabbit, `http://www.ncbi.nlm.nih.gov/nuccore/AAGW00000000.2`), *orcinus orca*

(orca, http://www.ncbi.nlm.nih.gov/nuccore/ANOL00000000.2), *monodelphis domestica* (opossum, http://www.ncbi.nlm.nih.gov/nuccore/AAFR00000000.3), *ornithorhynchus anatinus* (platypus, http://www.ncbi.nlm.nih.gov/nuccore/AAPN00000000.1), *crocodylus porosus* (crocodile, http://www.ncbi.nlm.nih.gov/nuccore/JRXG00000000.1), *danio rerio* (zebrafish, http://www.ncbi.nlm.nih.gov/nuccore/CABZ00000000.1). The results are displayed in Figure 3.6 in the main text.

Between-species analysis

Given larger diversity between gene segments in TRBV than IGHV, we measured diversity between gene segments in humans and dogs. The set of nucleotide sequences used for humans were the same as those used previously in the human within-species analysis. The set of nucleotide sequences used for dogs are curated at http://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=dog&group=IGHV. Again, only functional alleles were utilized in the study. Because identifying orthologous IGHV/TRBV genes in two species is very challenging, for each gene in the human reference, we computed the average alignment score to all other genes in the dog reference. For each gene in the human reference we could have used the alignment score to the closest aligning gene in the dog reference, but this might underestimate the amount of true gene diversity. Taking averages, we computed a mean score of –91 for IGHV and –306 for TRBV.

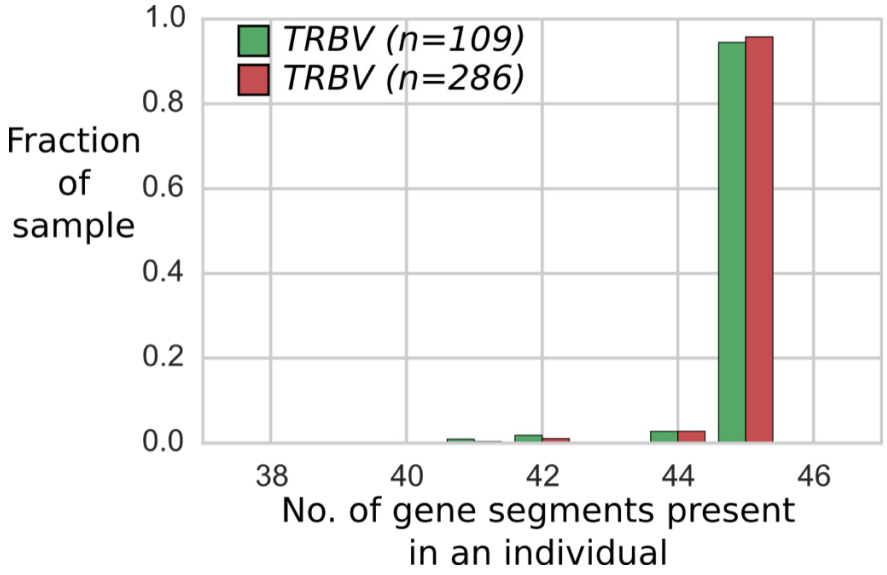


Figure B.9: The distribution of TRBV segments present in the sample of 109 individuals (from blood and saliva DNA; green) does not differ markedly from the distribution in extended sample of 286 individuals (blood, saliva, and cell line DNA; red).

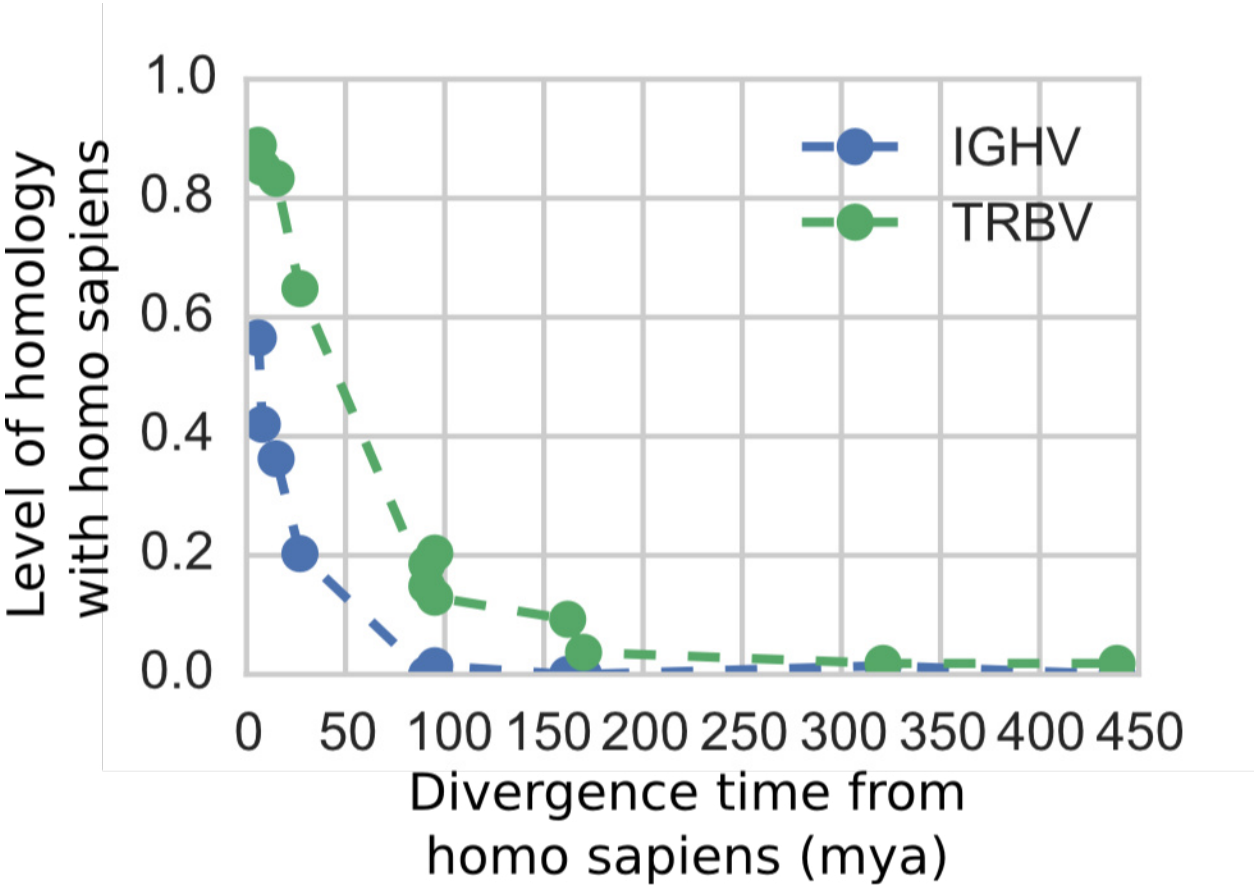


Figure B.10: The level of homology between homo sapiens segments and segments in other species plotted by the estimated divergence time between the species. The level of homology is defined as the fraction of homo sapiens segments that are more similar to segments in another species than another homo sapiens segment. Species are the same as in Figure 6. Divergence times are median estimates as reported by TimeTree.org.

First column	Second column
Allele name	Frequency
TRBV10-1*01	0.634
TRBV10-1*02_ gt234E_ (P)	0.239
TRBV10-1*02	0.127
TRBV10-2*01	0.894
TRBV10-2*01_tc191YY (T)	0.106
TRBV10-3*01	0.39
TRBV10-3*02	0.462

Continued on next page

Table B.1 – continued from previous page

Allele name	Frequency
TRBV10-3*03	0.124
TRBV10-3*01_ ga118GE (T)	0.024
TRBV11-1*01	0.94
TRBV11-1*01_ ag85HR_ct98YY_ag142QR	0.06
TRBV11-2*01	0.662
TRBV11-2*03	0.338
TRBV11-3*01	0.885
TRBV11-3*02	0.115
TRBV12-5*01	0.551
TRBV12-5*01_ cg27HD	0.417
TRBV12-5*01_ ga154RQ	0.032
TRBV13*01	0.977
TRBV13*01_ ct78PS (T)	0.023
TRBV14*01	0.866
TRBV14*02	0.134
TRBV15*02	0.977
TRBV15*01	0.023
TRBV16*01	0.976
TRBV16*02	0.024
TRBV18*01	0.953
TRBV18*01_ ag75MV	0.047
TRBV19*01	0.875
TRBV19*01_ ag23PP (T)	0.125
TRBV2*01	1.0
TRBV20-1*01	0.444
TRBV20-1*02	0.422
TRBV20-1*05	0.125
TRBV20-1*02_ ga227LL	0.009
TRBV27*01	1.0
TRBV28*01	1.0
TRBV29-1*01	0.944
TRBV29-1*01_ ac246ML (T)	0.056
TRBV3-1*01	1.0
TRBV30*01	0.707
TRBV30*02	0.211
TRBV30*01_ct108R__ct204PS (T)	0.02
TRBV30*02_ ct108R__ct204PS (P,T)	0.02
TRBV30*01_ga33VM (T)	0.014
TRBV30*04	0.027

Continued on next page

Table B.1 – continued from previous page

Allele name	Frequency
TRBV4-1*01	0.991
TRBV4-1*01_ cg181PR	0.009
TRBV5-1*01	1.0
TRBV5-4*01	0.991
TRBV5-4*01_tg213YD (T)	0.009
TRBV5-5*02	0.67
TRBV5-5*01	0.33
TRBV5-6*01	0.948
TRBV5-6*01_ta205FY_ct236NN	0.019
TRBV5-6*01_ta205FY	0.014
TRBV5-6*01_tg244LW	0.009
TRBV5-6*01_gt118GV	0.009
TRBV5-8*01_ct236NN (T)	0.09
TRBV5-8*01_ct55AV (T)	0.142
TRBV5-8*01	0.711
TRBV5-8*01_ct55AV_ct236NN (T)	0.057
TRBV6-1*01	0.991
TRBV6-1*01_ag183ND	0.009
TRBV6-4*01	0.869
TRBV6-4*02	0.085
TRBV6-4*02_ga49RQ	0.019
TRBV6-4*02_gc49RP	0.019
TRBV6-4*02_ac51SR	0.009
TRBV6-5*01	1.0
TRBV6-6*01	0.64
TRBV6-6*02	0.327
TRBV6-6*03_gt216DY	0.009
TRBV6-6*01_ga31RH (T)	0.014
TRBV6-6*01_ca278SR (T)	0.009
TRBV6-8*01	0.962
TRBV6-8*01_ag250QR	0.038
TRBV6-9*01	0.787
TRBV6-9*01_ag263VV	0.213
TRBV7-2*02	0.584
TRBV7-2*01	0.416
TRBV7-3*01	0.933
TRBV7-3*01_gt255DY (T)	0.067
TRBV7-4*01	0.954
TRBV7-4*01_ga214RK	0.037

Continued on next page

Table B.1 – continued from previous page

Allele name	Frequency
TRBV7-4*01_ct240RC	0.009
TRBV7-6*01	1.0
TRBV7-7*01	1.0
TRBV7-8*01	0.967
TRBV7-8*02	0.024
TRBV7-8*01_tc258SP (T)	0.009
TRBV7-9*01	0.226
TRBV7-9*03	0.747
TRBV7-9*03_ga67CY_ct105R__ag111TA (P)	0.028
TRBV9*01	0.894
TRBV9*02	0.106

Table B.1: Relative TRBV allele frequencies called from our sample of 109 individuals. Alleles are listed only if they were called in two or more individuals. The putative novel alleles are named by the closest matching allele in the IMGT database followed by mutations separated by '_'. Each mutation is represented as reference base pairalternative base pairpositionreference amino acidalternative amino acid. Note that '_' can also correspond to a stop codon, but this will be indicated by a 'P' in parentheses. A 'T' in parentheses denotes that the reconstructed allele sequence was truncated in one individual or more.

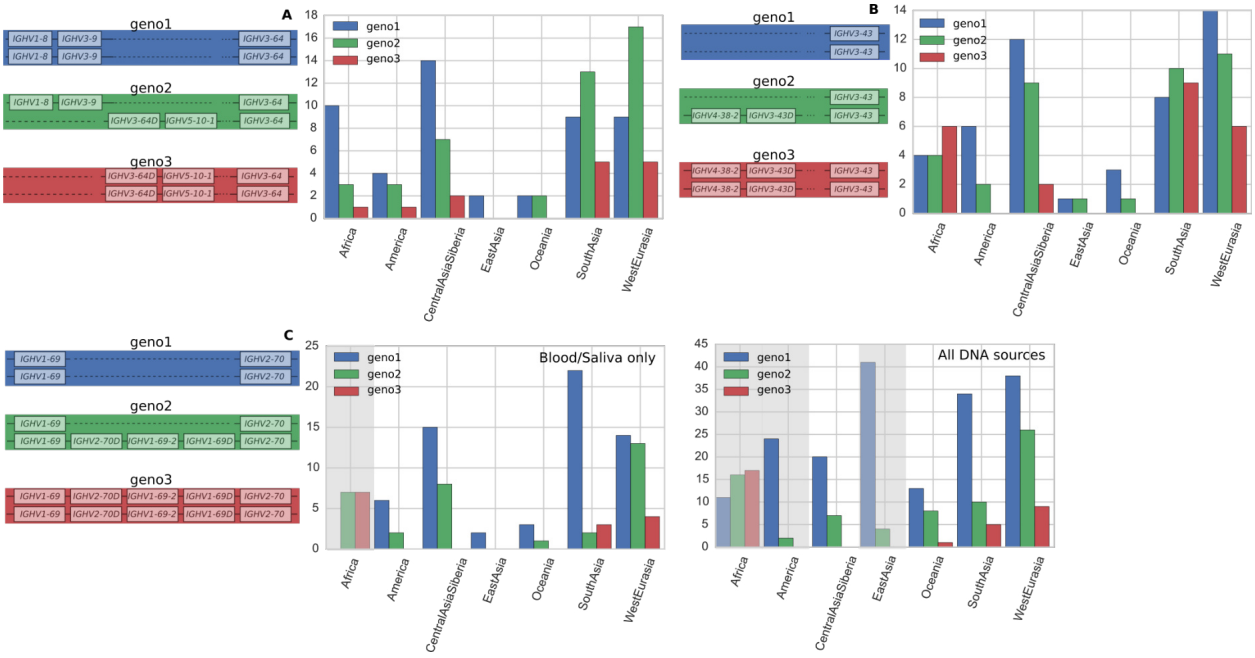


Figure B.11: Variant abundances for common IGHV polymorphisms within different geographical regions. (A) The IGHV polymorphism involving *IGHV1-8*, *IGHV3-9*, *IGVH5-10-1* and *IGHV3-64*. (B) The IGHV polymorphism involving *IGHV4-38-2* and *IGHV3-43*, *IGHV3-43D*. (C) The IGHV polymorphism involving *IGHV1-69*, *IGHV1-69-2*, and *IGHV2-70*. For all graphs, y-axis is the number of individuals. Figures are based on the 109 individuals with blood/saliva samples, except for the plot on the right in (C), where, because VDJ recombination is not believed to have a marked influence on copy number calls for the segments in the polymorphism, we use the full sample of 286 individuals from all DNA sources. Sample sizes for each region are 14 Africans, 31 West Eurasians, 23 Central Asians-Siberians, 2 East Asians, 27 South Asians, 4 Oceanians, 8 Native Americans. Grey shading indicates the distribution within a region is significantly different from the global distribution at the 0.01 level in a chi-squared goodness-of-fit test. Our data inform the copy number of these genes, while the genomic configuration is our best estimate based on previous studies [121, 87, 16, 10, 101, 100, 17, 74] (see Figure 3.2).

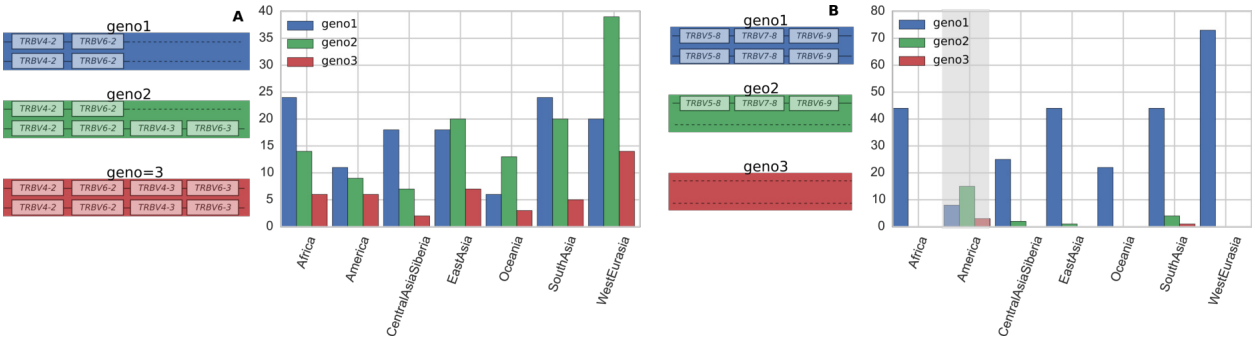


Figure B.12: Variant abundances for common TRBV copy number polymorphisms within different geographical regions. (A) The TRBV polymorphism involving *TRBV4-2*, *TRBV4-3*, *TRBV6-2*, and *TRBV6-3*. (B) The TRBV polymorphism involving *TRBV5-8*, *TRBV7-8*, and *TRBV6-9*. For both graphs, y-axis is the number of individuals. Both plots are based on the full sample of 286 individuals with 73 West Eurasians, 27 Central Asians-Siberians, 45 East Asians, 49 South Asians, and 22 Oceanians. Grey shading indicates the distribution within a region is significantly different from the global distribution at the 0.01 level in a chi-squared goodness-of-fit test. Our data informs the copy number of these genes, while the genomic configuration is our best estimate based on previous studies (see Figure 3.3).

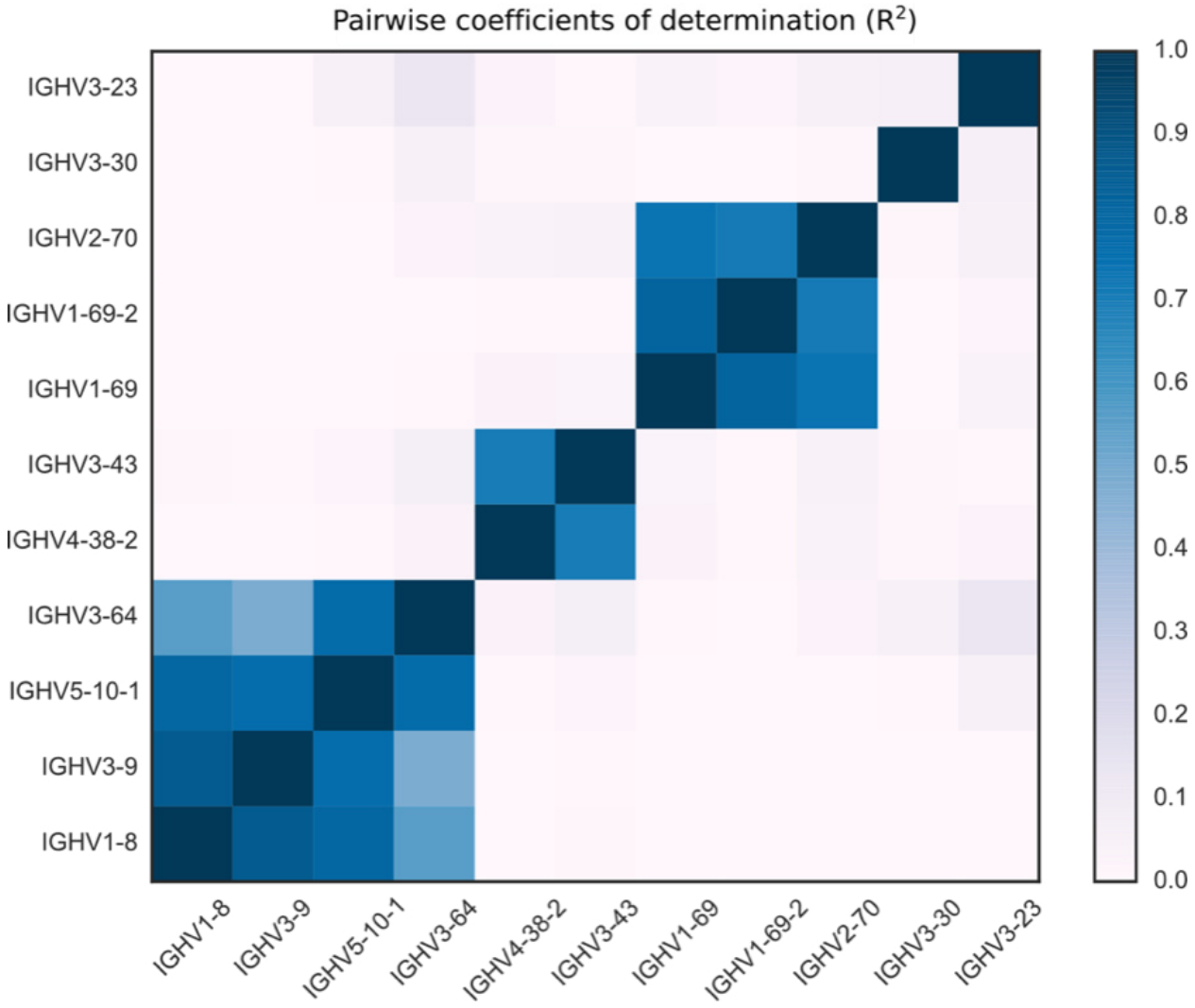


Figure B.13: Pairwise coefficient of determination between segments that appear in any IGHV copy number polymorphism. For each pair of segments, we calculate the coefficient of determination (also known as R squared) between the scaled copy number estimates at different gene segments over the sample of 109 individuals. A value of 1 corresponds to perfect correlation, a value of 0 to no correlation. Segments that are not in the same polymorphism have values very close to zero. Note that '*IGHV3-23*' refers to *IGHV3-23*, *IGHV3-23D*, '*IGHV3-30*' to *IGHV3-30*, *IGHV3-30-3*, *IGHV3-30-5*, *IGHV3-33*, '*IGHV3-43*' to *IGHV3-43*, *IGHV3-43D*, '*IGHV3-64*' to *IGHV3-64*, *IGHV3-64D*, '*IGHV1-69*' to *IGHV1-69*, *IGHV1-69D*, and '*IGHV2-70*' to *IGHV2-70*, *IGHV2-70D*.

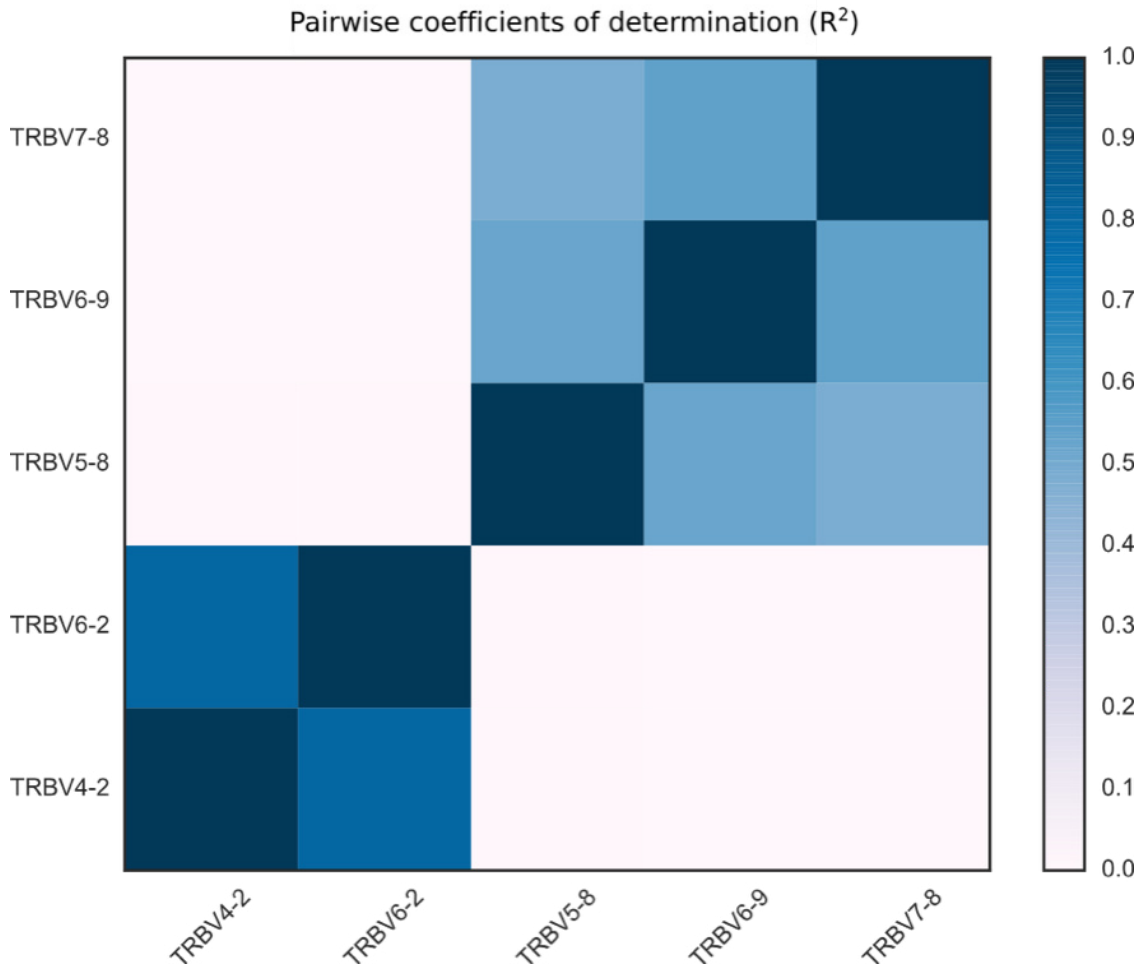


Figure B.14: Pairwise coefficient of determination between segments that appear in any TRBV copy number polymorphism. For each pair of segments, we calculate the coefficient of determination (also known as R squared) between the scaled copy number estimates at different gene segments over the full sample of 286 individuals. A value of 1 corresponds to perfect correlation, a value of 0 to no correlation. Segments that are not in the same polymorphism have values that are virtually zero. Note that '*TRBV4-2*' refers to *TRBV4-2*, *TRBV4-3* and '*TRBV6-2*' to *TRBV6-2*, *TRBV6-3*. Note that Figure 3.3 in the main text would suggest a correlation of 1 between the pairs of genes in *TRBV5-8*, *TRBV6-9*, *TRBV7-8*. However, as evidenced in Figure B.25, a large majority have two copies of all three genes, but the normalized coverage values have a large spread around 2. The lower correlation on the off-diagonal is likely due to the independent noisiness of the short-read whole-genome sequencing data.

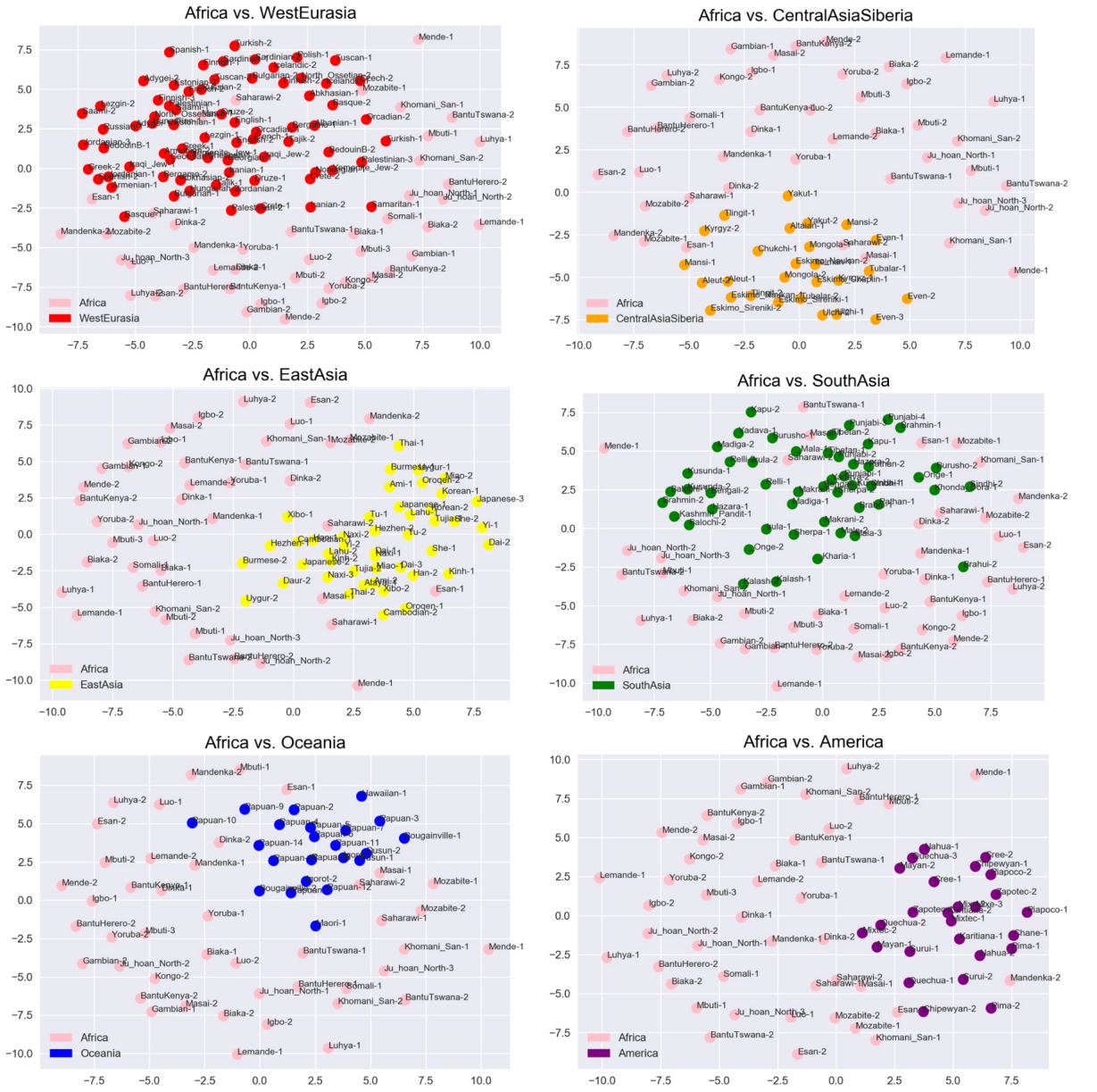


Figure B.15: Multidimensional scaling of 44 Africans and A) 73 West Eurasians, B) 27 Central Asians or Siberians, C) 45 East Asians, D) 49 South Asians, E) 22 Oceanians, and F) 26 Native Americans using inferred TRBV haplotypes. Data used includes all DNA source types in the SGDP dataset. The metric used for scaling was based on the Euclidean distance between the set of alleles of each individual. Individuals are labeled by an abbreviated version of their sample ID in the SGDP dataset.

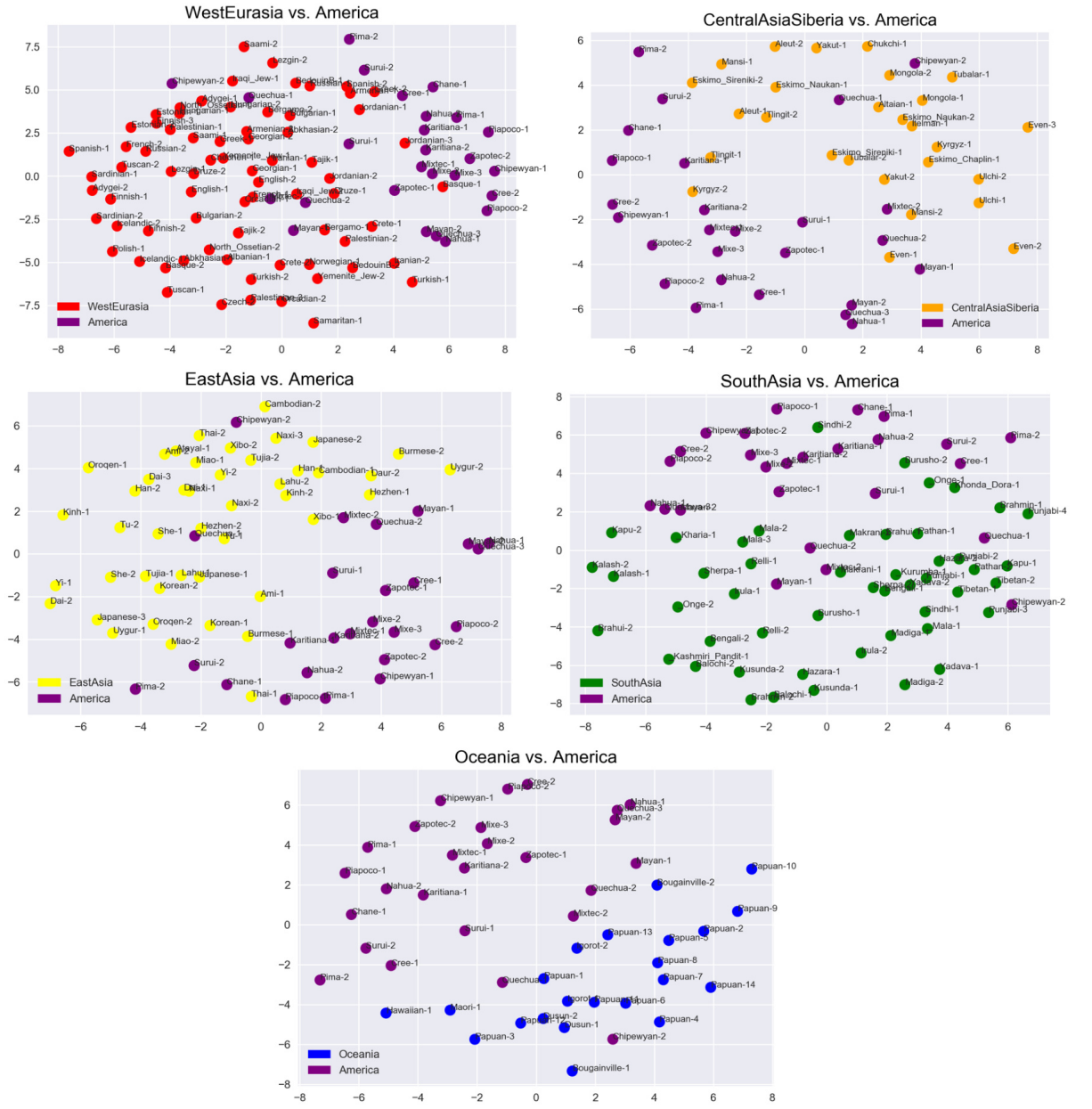


Figure B.16: Multidimensional scaling of 26 Native Americans and A) 73 West Eurasians, B) 27 Central Asians or Siberians, C) 45 East Asians, D) 49 South Asians, and E) 22 Oceanians using inferred TRBV haplotypes. For comparison of Native Americans and Africans, refer to panel F) of Fig. S9. Data used includes all DNA source types in the SGDP dataset. The metric used for scaling was based on the Euclidean distance between the set of alleles of each individual. Individuals are labeled by an abbreviated version of their sample ID in the SGDP dataset.

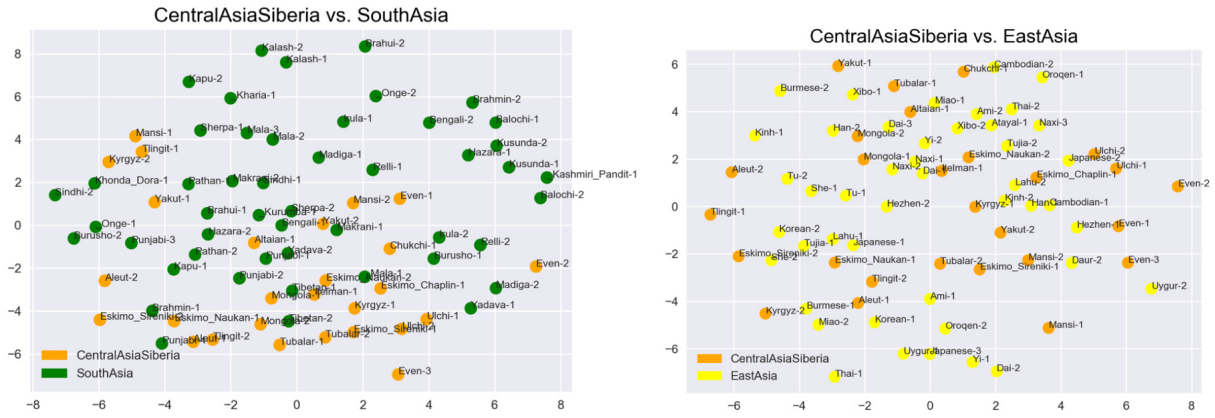


Figure B.17: Multidimensional scaling of 27 individuals from Central Asia–Siberia and A) 49 individuals from South Asia and B) 45 individuals from East Asia using inferred TRBV haplotypes. Data used includes all DNA source types in the SGDP dataset. The metric used for scaling was based on the Euclidean distance between the set of alleles of each individual. Individuals are labeled by an abbreviated version of their sample ID in the SGDP dataset.

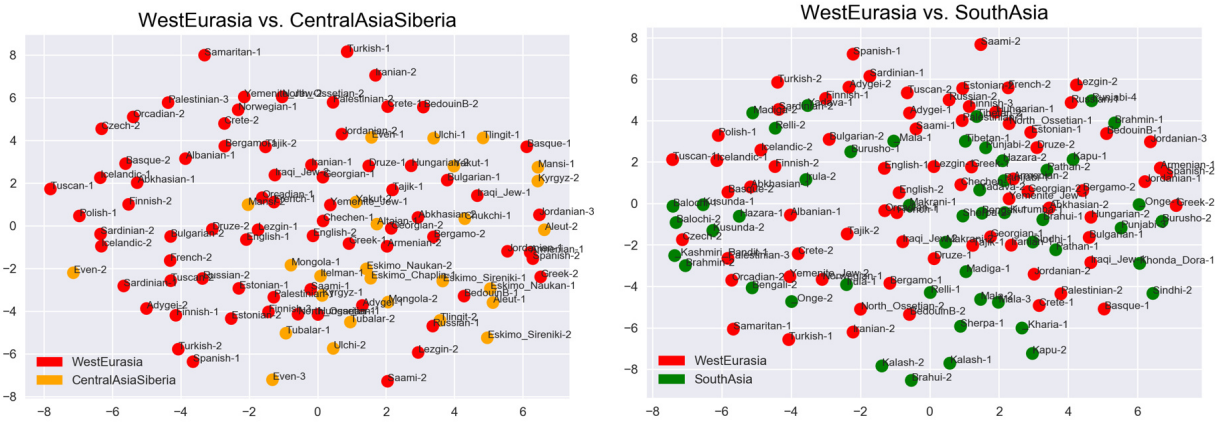


Figure B.18: Multidimensional scaling of 73 individuals from West Eurasia and A) 27 individuals from Central Asia-Siberia and B) 49 individuals from South Asia using inferred TRBV haplotypes. Data used includes all DNA source types in the SGDP dataset. The metric used for scaling was based on the Euclidean distance between the set of alleles of each individual. Individuals are labeled by an abbreviated version of their sample ID in the SGDP dataset.

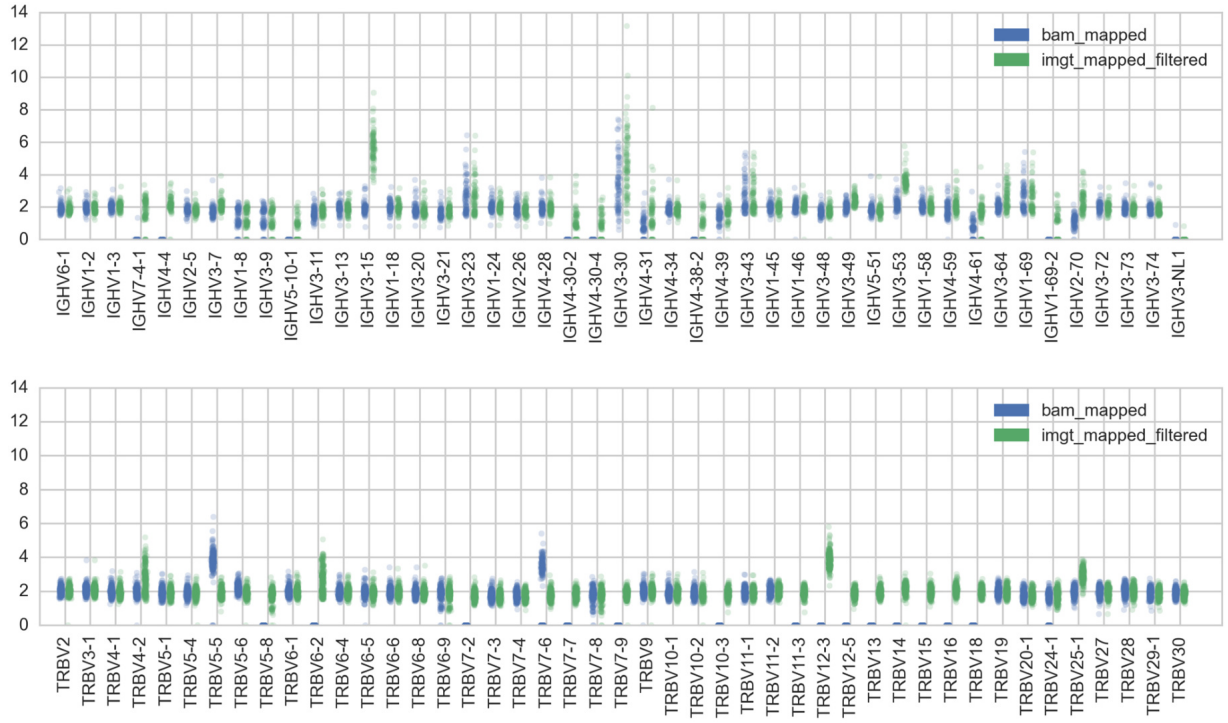


Figure B.19: Differences in copy number estimation using reads mapped to GRCh37 versus reads mapped to IMGT alleles. IGHV and TRBV gene segments that are not in GRCh37 assembly are systematically missing from the reads collected via read mapping to GRCh37 (blue, 'bam_mapped', procedure (i)) compared to reads mapped to functional IMGT alleles (green, 'imgt_mapped_filtered', procedure (ii) with additional filtering).

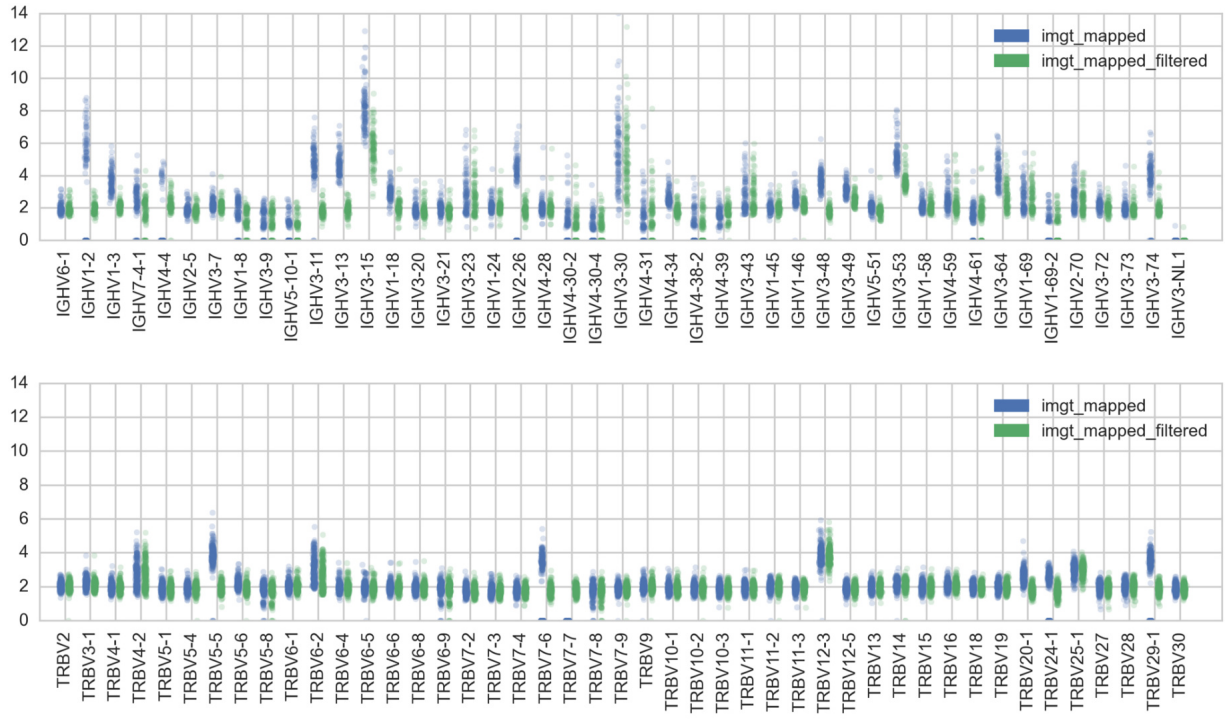


Figure B.20: Directly calculating copy number from IMGT-mapped reads (blue, 'imgt_mapped', procedure (ii)) leads to overestimates of copy number. This is most likely due to reads from similar pseudogenes and orphon genes being erroneously mapped to a functional gene. These overestimates are reduced when further filtering procedures are applied (green, 'imgt_mapped_filtered').

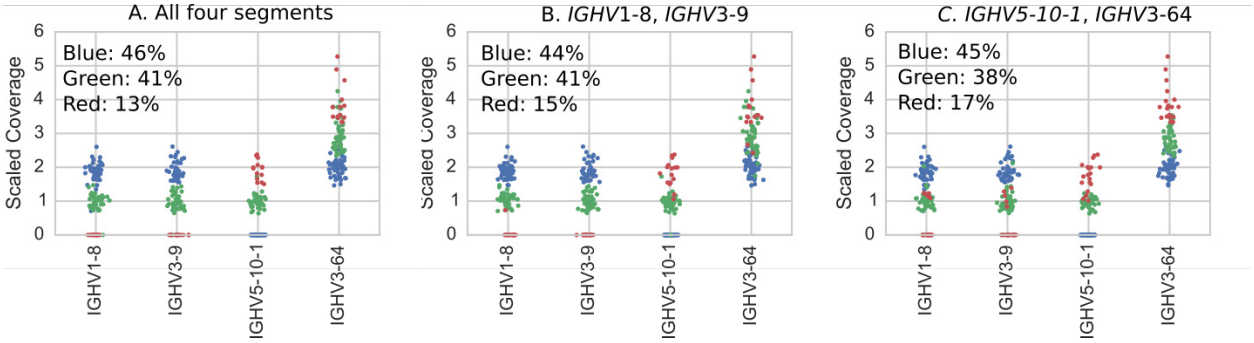


Figure B.21: Individuals (dots) colored according to results of hierarchical clustering of copy number estimates (as described in Supplementary Text) for *IGHV1-8*, *IGHV3-9*, *IGHV5-10-1*, and *IGHV3-64*, *IGHV3-64D* gene segments. Colors correspond to the variants in Figure B.11 and the plot titles describes the genes used in the clustering. (A) shows clustering using copy number estimates from all four segments. These are the results we report in Figure 3.2. (B) and (C) shows clustering using copy number estimates of only subsets of the genes comprising the polymorphism. Note that even when only subsets of the genes are used, the clustering is still clear in the genes that were not used in the clustering but which are part of the polymorphism.

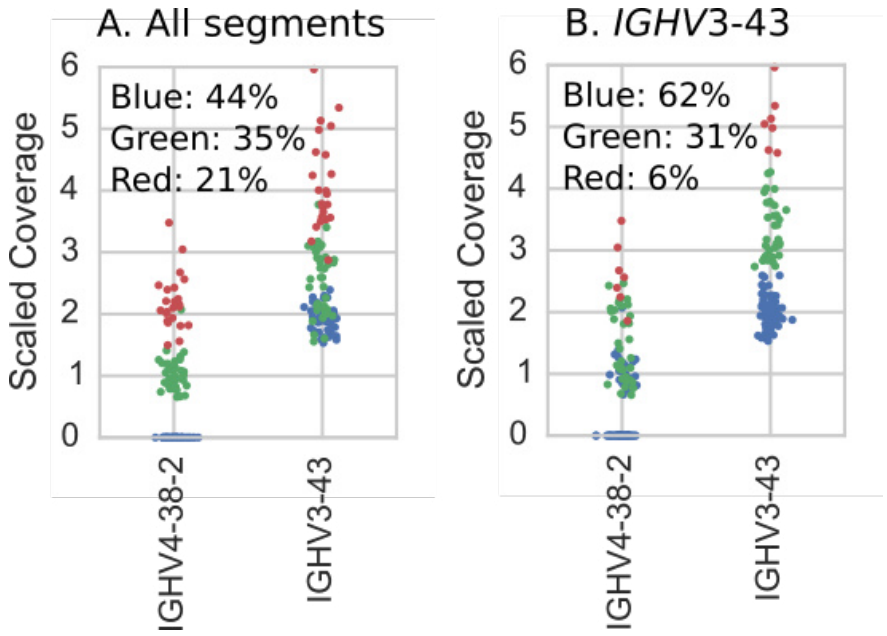


Figure B.22: Individuals (dots) colored according to results of hierarchical clustering of copy number estimates (as described in Supplementary Text) for *IGHV4-38-2* and *IGHV3-43*, *IGHV3-43D* gene segments. Colors correspond to the variants in Figure B.11 and the plot titles describe the genes used in the clustering. (A) shows clustering using copy number estimates from both segments. These are the results we report in Figure 3.2. (B) shows clustering using just the copy number estimates for *IGHV3-43*, *IGHV3-43D*. Note that when only one operationally distinguishable gene is used for performing clustering, the clusters are determined by hard numerical thresholds.

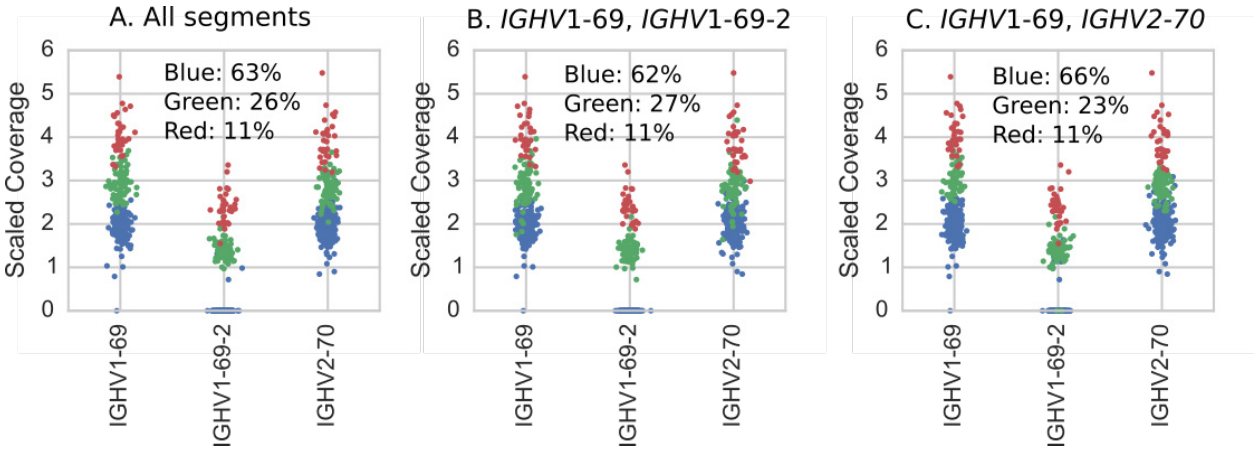


Figure B.23: All 286 individuals (dots) colored according to results of hierarchical clustering of copy number estimates (as described in Supplementary Text) for *IGHV1-69*, *IGHV1-69D*, *IGHV1-69-2*, and *IGHV2-70*, *IGHV2-70D* gene segments. Colors correspond to the variants in Figure B.11 and the plot titles describe the genes used in the clustering. (A) shows clustering using copy number estimates from all segments. These are the results we report in Figure 3.2. (B) and (C) shows clustering using copy number estimates of only subsets of the genes comprising the polymorphism. Note that even when only subsets of the genes are used, the clustering is still clear in the genes that were not used in the clustering but which are part of the polymorphism.

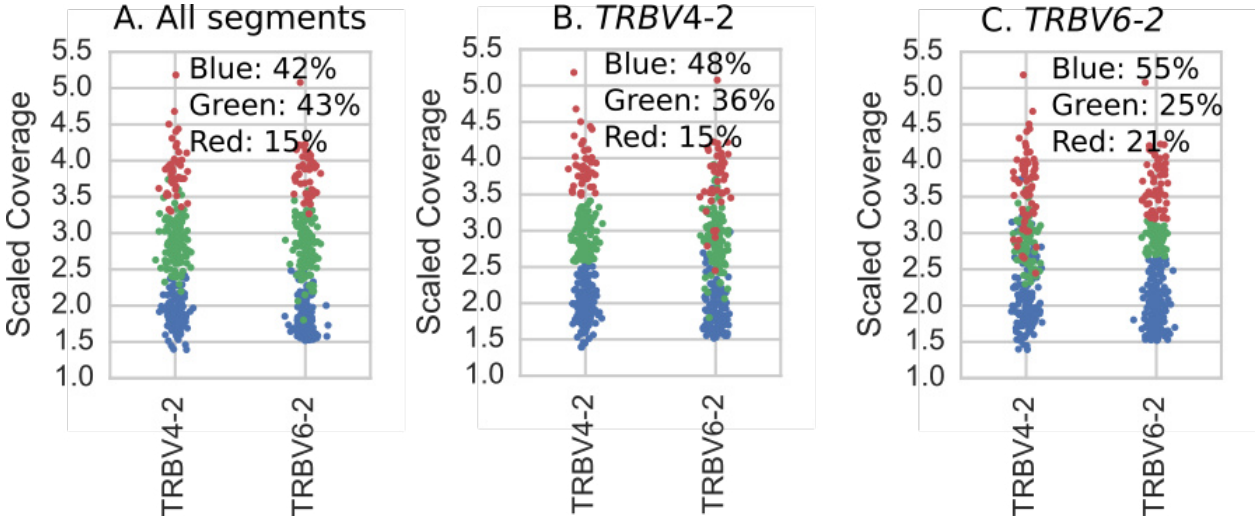


Figure B.24: All 286 individuals (dots) colored according to results of hierarchical clustering of copy number estimates (as described in Supplementary Text) for *TRBV4-2*, *TRBV4-3* and *TRBV6-2*, *TRBV6-3* gene segments. Colors correspond to the variants in Figure B.12 and the plot titles describe the genes used in the clustering. (A) shows clustering using copy number estimates from both segments. These are the results we report in Figure 3.3. (B) and (C) show clustering using just the copy number estimates for each operationally distinguishable gene individually. Note that as with Figure B.22 (*IGHV4-38-2*, *IGHV3-43*), when only one operationally distinguishable gene is used for performing clustering, the clusters are determined by hard numerical thresholds.

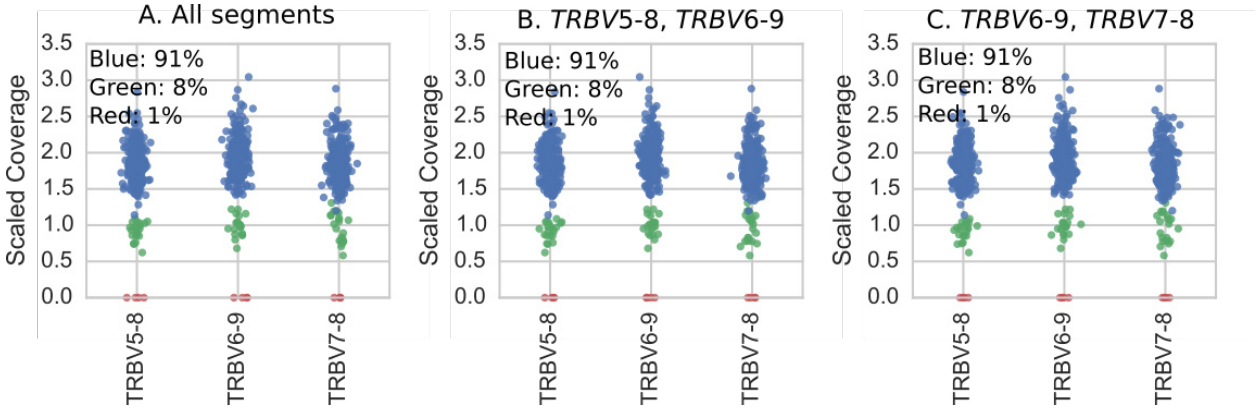


Figure B.25: All 286 individuals (dots) colored according to results of hierarchical clustering of copy number estimates (as described in Supplementary Text) for *TRBV5-8*, *TRBV6-9*, and *TRBV7-8* gene segments. Colors correspond to the variants in Figure B.12 and the plot titles describe the genes used in the clustering. (A) shows clustering using copy number estimates from all segments. These are the results we report in Figure 3.3. (B) and (C) shows clustering using copy number estimates of only subsets of the genes comprising the polymorphism. Note that even when only subsets of the genes are used, the clustering is still clear in the genes that were not used in the clustering but which are part of the polymorphism.

Allele name	Allele frequency
IGHV1-18*01	0.824
IGHV1-18*04	0.167
IGHV1-18*01_ag168ND	0.009
IGHV1-24*01	1.0
IGHV1-45*02	0.839
IGHV1-45*02_ga123GR	0.161
IGHV1-58*01	0.465
IGHV1-58*02	0.498
IGHV1-58*02_gt57VF	0.037
IGHV2-26*01	0.821
IGHV2-26*01_ct257NN_ct294RW	0.113
IGHV2-26*01_ct257NN	0.066
IGHV3-20*01	0.561
IGHV3-20*01_ct282HY	0.198
IGHV3-20*01_gt64CF	0.231
IGHV3-20*01_ag88DG_ct282HY	0.009
IGHV3-72*01	0.991
IGHV3-72*01_tc170SS	0.009
IGHV3-73*01	0.502
IGHV3-73*02	0.479
IGHV3-73*01_ag55KR	0.018
IGHV3-74*01	0.986
IGHV3-74*02	0.014
IGHV5-51*01	0.824
IGHV5-51*03	0.157
IGHV5-51*01_ga112RH	0.019
IGHV6-1*01	0.963
IGHV6-1*01_ct207R_ (P)	0.037

Table B.2: Relative IGHV allele frequencies called from our sample of 109 individuals. Alleles are listed only if they were called in two or more individuals. The putative novel alleles are named by the closest matching allele in the IMGT database followed by mutations separated by '_'. Each mutation is represented as reference base pairalternative base pairpositionreference amino acidalternative amino acid. For example, allele '*IGHV1-18*01_ag168ND*' denotes an allele whose sequence is that of IMGT allele *IGHV1-18*01*, but with a 'g' at position 168 rather than an 'a'. The two letters following the position correspond to the amino acid with the reference base pair 'a' and the amino acid with the mutation 'g', respectively. Note that '_' can also correspond to a stop codon, but this will be indicated by a 'P' in parentheses. A 'T' in parentheses denotes that the reconstructed allele sequence was truncated in one individual or more.

	Africans	WE	CAS	EA	SA	Oceanians
WE	0.126					
CAS	0.165	0.0453				
EA	0.195	0.0416	0.0181			
SA	0.0899	0.0082	0.0495	0.0635		
Oceanians	0.152	0.0402	0.0384	0.0188	0.0489	
NA	0.15	0.1336	0.129	0.17	0.121	0.202

Table B.3: FST results using SNP information for the two-copy TRBV gene segments of 289 individuals for the seven defined regions. FST computations were done using Genepop for the seven geographic regions: 44 Africans, 73 West Eurasians (WE), 27 Central Asian-Siberians (CAS), 45 East Asians (EA), 49 South Asians (SA), 22 Oceanians, and 26 Native Americans (NA). Overall estimate of FST is 0.0955 (prior estimates are 0.14 using the entire TRBV region [110].

Allele name	A	WE	CAS	EA	SA	O	NA	Total
IGHV1-45*02_ga123GR	0.04	0.08	0.26	0.5	0.17	0.25	0.25	0.16
IGHV2-26*01_ct257NN_ct294RW	0.04	0.05	0.24	0	0.11	0	0.19	0.11
IGHV2-26*01_ct257NN	0.35	0.02	0.09	0	0	0.14	0	0.07
IGHV3-20*01_ct282HY	0.52	0.11	0.09	0	0.24	0.25	0.19	0.2
IGHV3-20*01_gt64CF	0.11	0.33	0.26	0	0.2	0	0.25	0.23
TRBV10-1*02_gt234E_ (P)	0.19	0.28	0.24	0.33	0.15	0.62	0.25	0.24
TRBV10-2*01_tc191YY (T)	0.11	0.13	0.02	0	0.21	0	0	0.11
TRBV11-1*01_ ag85HR_ct98YY_ag142QR	0.07	0.16	0	0	0.02	0	0	0.06
TRBV12-5*01_cg27HD	0.11	0.4	0.57	0.25	0.5	0.88	0.12	0.42
TRBV18*01_ag75MV	0.36	0.02	0	0	0	0	0	0.05
TRBV19*01_ag23PP (T)	0.04	0.11	0.18	0.25	0.09	0	0.33	0.12
TRBV29-1*01_ac246ML (T)	0	0	0.16	0.25	0	0	0.25	0.06
TRBV5-8*01_ct236NN (T)	0	0.08	0.07	0.25	0.15	0	0.17	0.09
TRBV5-8*01_ct55AV (T)	0.39	0.08	0.07	0.25	0.15	0	0.17	0.14
TRBV5-8*01_ct55AV_ct236NN (T)	0.07	0.03	0	0	0.12	0	0.17	0.06
TRBV6-9*01_ag263VV	0.5	0.13	0.09	0.25	0.27	0	0.33	0.21
TRBV7-3*01_gt255DY (T)	0	0	0	0.25	0.24	0	0	0.07

Table B.4: Allele frequencies in our sample of 109 individuals (218 haplotypes) for putatively novel alleles that appear at least 10 times in the sample for IGHV (top) and TRBV (bottom), by geographic region (14 Africans (A), 31 West Eurasians (WE), 23 Asians–Siberians (CAS), 2 East Asians (EA), 27 South Asians (SA), 4 Oceanians (O), 8 Native Americans (NA)) and all regions (Total). The putative novel alleles are named by the closest matching allele in the IMGT database followed by mutations separated by '_'. Each mutation is represented as reference base pairalternative base pairpositionreference amino acidalternative amino acid. Note that '_' can also correspond to a stop codon, but this will be indicated by a 'P' in parentheses. A 'T' in parentheses denotes that the reconstructed allele sequence was truncated in one individual or more.

Variant name	Variant type	Region	#	Regional freq	Global freq
IGHV1-18_g168	SNV	Africa	2	0.07	0.01
IGHV1-58_t57	SNV	Africa	8	0.29	0.04
IGHV2-26_a118	SNV	Africa	2	0.07	0.01
IGHV3-72_c170	SNV	Africa	2	0.07	0.01
IGHV3-74_t20	SNV	Africa	3	0.11	0.01
IGHV1-18*01_ag168ND	Allele	Africa	2	0.07	0.01
IGHV1-58*02_gt57VF	Allele	Africa	8	0.29	0.04
IGHV3-72*01_tc170SS	Allele	Africa	2	0.07	0.01
IGHV3-74*02	Allele	Africa	3	0.12	0.01
TRBV13_t78	SNV	CAS	5	0.11	0.02
TRBV20-1_a227	SNV	Africa	2	0.08	0.01
TRBV30_a33	SNV	CAS	2	0.08	0.01
TRBV4-1_g181	SNV	Africa	2	0.07	0.01
TRBV5-4_g213	SNV	Africa	2	0.08	0.01
TRBV5-6_t118	SNV	SouthAsia	2	0.04	0.01
TRBV5-6_a205	SNV	Africa	8	0.29	0.04
TRBV5-6_t236	SNV	Africa	4	0.14	0.02
TRBV6-1_g183	SNV	Africa	2	0.08	0.01
TRBV6-6_a31	SNV	Africa	3	0.12	0.01
TRBV6-6_t216	SNV	Africa	2	0.08	0.01
TRBV6-6_a278	SNV	SouthAsia	2	0.04	0.01
TRBV6-8_g250	SNV	Africa	8	0.42	0.04
TRBV7-4_t240	SNV	Africa	2	0.07	0.01
TRBV13*01_ct78PS (T)	Allele	CAS	5	0.11	0.02
TRBV20-1*02_ga227LL	Allele	Africa	2	0.08	0.01
TRBV30*01_ga33VM (T)	Allele	CAS	2	0.08	0.01
TRBV4-1*01_cg181PR	Allele	Africa	2	0.07	0.01
TRBV5-4*01_tg213YD (T)	Allele	Africa	2	0.08	0.01
TRBV5-6*01_ta205FY_ct236NN	Allele	Africa	4	0.15	0.02
TRBV5-6*01_ta205FY	Allele	Africa	3	0.11	0.01
TRBV5-6*01_gt118GV	Allele	SouthAsia	2	0.04	0.01
TRBV6-1*01_ag183ND	Allele	Africa	2	0.08	0.01
TRBV6-6*03_gt216DY	Allele	Africa	2	0.08	0.01
TRBV6-6*01_ga31RH (T)	Allele	Africa	3	0.12	0.01
TRBV6-6*01_ca278SR (T)	Allele	SouthAsia	2	0.04	0.01
TRBV6-8*01_ag250QR	Allele	Africa	8	0.42	0.04
TRBV7-4*01_ct240RC	Allele	Africa	2	0.07	0.01

Table B.5: SNVs and putative novel alleles in our sample of 218 haplotypes that are private to a geographic region (14 Africans, 31 West Eurasians, 23 Central Asians-Siberians (CAS), 2 East Asians, 27 South Asians, 4 Oceanians, 8 Native Americans). See Section B.7 for notation explanation.

Region	IGHV (n=109)	TRBV (n=286)
Africans	0.98	0
Native Americans	1	0.34
Central Asians or Siberians	0.98	0.08
East Asians	1	0
Oceanians	1	0
South Asians	0.99	0
West Eurasians	0.98	0.1

Table B.6: Table of probabilities that two individuals drawn at random from the same geographic region in our sample have different sets of IGHV or TRBV segments. Two sets are considered different if there is at least one operationally distinguishable segment that is present (in any number of copies) in one set but is absent in the other set.