UNIVERSITY OF CALIFORNIA SAN DIEGO

Data-driven insights from immunogenomics, metabolomics, and clinical mental health data

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Machine Learning and Data Science)

by

Vinnu Bhardwaj

Committee in charge:

Professor Ramesh Rao, Chair
Professor Massimo Franceschetti, Co-Chair
Professor Dewleen Baker
Professor Mohit Jain
Professor Siavash Mirarab
Professor Pavel Pevzner

2020

The Dissertation of Vinnu Bhardwaj is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-chair

_____

Chair

University of California San Diego

2020

DEDICATION

To my family, whose love and support prepared me for the journey through graduate school.

To Sonal, who made this journey delightful.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I want to thank my advisors Ramesh and Massimo whose guidance all along was invaluable. They supported and even encouraged me to take risks and go out of my comfort zone and take up and pursue projects I knew nothing about, especially in the field of computational biology. I learnt a great deal from them on how to collaborate with people from different disciplines.

I have been very fortunate to have worked alongside so many talented researchers over the course of my PhD. A very special thanks to Yana Safonova and Pavel Pevzner. They introduced me to the field of immunogenomics and guided me in the right directions while I was working on the immunogenomics projects. For the mental health project, I worked with Dewleen and Abigail, whose expertise on the subject helped me tremendously. In the final year of my PhD, I worked in collaboration with Jain Lab at the UCSD School of Medicine. I would like to thank Mohit, Jeramie, Tao, and Shriram for introducing me to the field of metabolomics and their support while I was working there. Finally, I am grateful to Siavash, who agreed to serve on my committee and provide valuable feedback during the oral exams.

Chapter 2 is adapted from Bhardwaj, V., Franceschetti, M., Rao, R., Pevzner, P. A., & Safonova, Y. (2020). Automated analysis of immunosequencing datasets reveals novel immunoglobulin D genes across diverse species. PLOS Computational Biology, 16(4), e1007837. The dissertation author was the primary author of this paper.

Chapter 3 is being prepared for submission for publication of the material. The material is co-authored with members from the Jain Lab, Nallamshetty S., and Rao. R. The dissertation author was one of the primary authors of this material.

Chapter 4 is adapted from Bhardwaj, V., Angkaw, A. C., Franceschetti, M., Rao, R., & Baker, D. G. (2019). Direct and indirect relationships among posttraumatic stress disorder, depression, hostility, anger, and verbal and physical aggression in returning veterans. Aggressive behavior, 45(4), 417-426. The dissertation author was the primary author of this paper.

Vita

2013 Bachelor of Engineering, PEC University of Technology, Chandigarh, India.

2015 Master of Engineering, Indian Institute of Science, Bengaluru, India.

2020 Doctor of Philosophy, University of California San Diego, San Diego, USA.

Publications

Bhardwaj V, Franceschetti M, Rao R, Pevzner PA, Safonova Y (2020) Automated analysis of immunosequencing datasets reveals novel immunoglobulin D genes across diverse species. PLOS Computational Biology.

Bhardwaj V, Pevzner P, Rashtchian C, Safonova Y. Trace reconstruction problems in computational biology, submitted for publication in IEEE Transactions on Information Theory.

Bhardwaj V, Angkaw AC, Franceschetti M, Rao R, Baker DG (2019). Direct and indirect relationships among posttraumatic stress disorder, depression, hostility, anger, and verbal and physical aggression in returning veterans. Aggressive Behavior.

ABSTRACT OF THE DISSERTATION

Data-driven insights from immunogenomics, metabolomics, and clinical mental health data

by

Vinnu Bhardwaj

Doctor of Philosophy in Electrical Engineering (Machine Learning and Data Science)

University of California San Diego, 2020

Professor Ramesh Rao, Chair

Professor Massimo Franceschetti, Co-Chair

In recent years, we have seen the rise of technologies with unprecedented abilities to query the status of different biological systems that generate enormous amounts of data. While these data hold the promise to unravel new insights and better understanding of the working of the human biological systems, this access to enormous amounts of data is frequently accompanied by computational and algorithmic challenges. This dissertation answers questions related to some of these systems with the help of data – in some cases by developing new algorithms and computational tools, and in others using statistical and exploratory data analyses approaches.

Chapter 1

Introduction

High-throughput technologies have enabled the creation of numerous experiments probing DNA, RNA, and metabolites. The vast amount of untargeted data generated by such technologies can enable inferring novel insights about biological mechanisms in the human body by way of exploratory data analyses. Frequently, with the rise of new technologies and new types of data comes the need of new computational tools. This dissertation answers specific questions about three areas in human health – immunology, metabolism, and mental health - in the next three chapters, respectively.

Chapter 2 of this dissertation presents a novel computational method for the inference of immunoglobulin genes using immunosequencing data. Antibodies (immunoglobulins) provide specific binding to an enormous range of antigens and represent a key component of the adaptive immune system. Immunosequencing has emerged as a method of choice for generating millions of reads that sample antibody repertoires and provides insights into monitoring immune response to disease and vaccination. Immunoglobulin genes are formed through V(D)J recombination, which joins the variable (V), diversity (D), and joining (J) germline genes. Since variations in germline genes have been linked to various diseases, personalized immunogenomics focuses on finding alleles

of germline genes across various patients. Efforts such as the 1000 Genomes Project have led to only limited progress towards inferring the population-wide census of germline immunoglobulin genes because of the difficulty in assembling the highly repetitive immunoglobulin loci from whole genome sequencing data. Chapter 2 describes a novel computational method for the inference of D genes from immunosequencing data.

Chapter 3 of this dissertation studies the effects of fasting on human health using data gathered by nontargeted liquid chromatography mass spectroscopy (LCMS). A growing body of evidence suggests that factors related to meal timing, frequency, and caloric content significantly impact health and longevity. Long-term caloric restriction and intermittent periods of fasting have been linked to a wide array of benefits in humans. The molecular mechanisms that underlie the long-term benefits of fasting are thought to be linked at least in part to an orchestrated shift in fuel metabolism in non-hepatic tissues. Chapter 3 presents part of our ongoing work on the systematic analysis of potentially protective factors that increase with fasting.

Chapter 4 of this dissertation studies the influence of war on the mental health of post-combat veterans using data from a sample of veterans returning from Iraq or Afghanistan. Hostility, anger, and aggression are conceptually related but unique constructs found to occur more often among veterans with posttraumatic stress disorder (PTSD) than among civilians or veterans without PTSD. The pathways between PTSD, depression, hostility, anger, and aggression have not been comprehensively characterized. Direct and indirect relationships among PTSD, depression, hostility, anger, and four types of aggression - verbal, and physical toward self, others, and objects – are investigated in Chapter 4.

Chapter 2

Personalized inference of IGHD genes using immunosequencing data

2.1 Introduction

Antibodies provide specific binding to an enormous range of antigens and represent a key component of the adaptive immune system [1]. The *antibody repertoire* is generated by *somatic recombination* of the V (*variable*), D (*diversity*), and J (*joining*) germline genes by a process known as V(D)J recombination. During this process, the germline V, D, and J genes are randomly selected, and the gene ends are randomly trimmed and joined together along with some random insertions between the trimmed genes, leading to a huge number of unique recombined sequences. The specificity of an antibody is largely defined by the recombination site referred to as the *third complementarity determining region* (*CDR3*) [2].

Immunosequencing helps in monitoring immune response to disease and vaccination by generating millions of reads that sample antibody repertoires [3]. Information about all germline immunoglobulin genes specific to the *individual* is a prerequisite for analyzing immunosequencing (*Rep-Seq*) data. However, most previous immunogenomics studies have relied on the *population-*

level germline genes. As the set of known germline genes is incomplete (particularly for non-Europeans) and contains alleles that resulted from sequencing and annotation errors [4, 5], studies based on population-level germline genes can lead to incorrect results. Moreover, it is difficult to find which known allele(s) is present in a specific individual since the widespread practice of aligning each read to its closest germline gene results in high error rates [5]. Using population-level germline genes rather than individual germline genes can thus make it difficult to analyze *somatic hypermutations* (*SHM*) and clonal development of antibody repertoires [6-8].

Identifying individual germline genes (i.e., *personalized immunogenomics*) is important since variations in germline genes have been linked to various diseases [9], differential response to infection, vaccination, and drugs [10, 11], aging [12], and disease susceptibility [9, 13, 14]. There still exist unknown human allelic variants and the International ImMunoGeneTics (IMGT) database [15] is incomplete even for well-studied human germline genes [16]. The germline genes for less studied albeit immunologically important model organisms remain largely unknown [17, 18]. Assembling the highly repetitive immunoglobulin loci from whole genome sequencing data is difficult [19] and efforts such as the 1000 Genomes Project have led to only limited progress towards inferring the population-wide census of germline immunoglobulin genes [19-21].

Although the personalized immunogenomics approach was first proposed by [22], the manual analysis in this study did not result in a software tool for inferring germline genes. Gadala-Maria et al. [23] developed the TIgGER algorithm for inferring germline genes and used it to discover novel alleles of V genes. The challenge of *de novo* reconstruction of V and J genes was further addressed by Corcoran et al. [24], Zhang et al. [25], Ralph and Matsen [5], and Gadala-Maria et. al. [26]. However, as Ralph and Matsen [5] commented, the more challenging task of *de novo* reconstruction of D genes remained elusive.

The sequences encoded by D genes play important roles in B cell development, antigen binding site diversity, and antibody production [27]. Safonova and Pevzner [28] recently developed the IgScout algorithm for *de novo* inference of D genes using immunosequencing data. Unlike

algorithms for de novo inference of V and J genes [23, 24], it does not rely on alignments against closest germline genes that might lead to erroneous inferences [29, 30]. Instead, IgScout uses the observation that the most abundant *k*-mers in CDR3s arise from D genes (a *k-mer* refers to a string of length *k*). However, IgScout lacks a probabilistic model and has limitations with respect to inferring short D genes and D genes that share substrings with other D genes. It relies on the knowledge of *k* such that each *k*-mer occurs in a single D gene (information that is often unavailable) and uses those *k*-mers as *seeds* in its *seed extension* procedure. However, if a *k*-mer seed occurs in multiple D genes, IgScout might miss some D genes altogether and sometimes even produce inaccurate results. To bypass this problem, IgScout attempts to select large *k* to guarantee that each *k*-mer occurs in a single D gene (e.g., k=15 for human D genes). However, using long *k*-mers as seeds results in missing D genes that are shorter than those *k*-mers. Thus, for species with limited information about the range of D gene lengths, IgScout is bound to make errors.

Our MINING-D algorithm uses a probabilistic model and addresses above limitations of IgScout. We applied MINING-D to nearly 600 publicly available Rep-seq datasets from humans, mice, camels, rhesus macaques, rats, and rabbits. In total, MINING-D inferred 13, 6, 4, 8, 12, and 15 novel D genes using human, mouse, rat, macaque, camel, and rabbit datasets, respectively. We validated 25 out of these 58 novel D genes - 2, 1, 3, 8, 8, 3 D genes for human, mouse, rat, macaque, camel, and rabbit datasets, respectively - using Whole Genome Sequencing data. We further analyzed the usage of D genes in diverse Rep-seq datasets to analyze potential associations between the usage of a D gene and an environment, *i.e.,* a health condition, a tissue, or a cell type.

2.2 Methods

2.2.1 Probabilistic model of CDR3 generation.

The transformation of a D gene (a *seed string*) to a CDR3 (a *modified string)* can be modeled by the following probabilistic model. The seed string s is *trimmed* at two randomly chosen locations p and q ($p+q \leq |s|$, where $|.|$ denotes the length of a string) such that the first p and the last q symbols

of s are removed (Figure 2.1a). The resulting string is extended on the left and on the right by randomly generated strings e_l and e_r of randomly selected lengths l_l and l_r respectively. The resulting string is further extended on the left by a randomly chosen string v_l from a set of strings V_{cdr3} and on the right by a randomly chosen string j_r from a set of strings J_{cdr3} to form a *modified string* c.

The seed string s in the above model corresponds to a D gene, the strings e_l and e_r correspond to the random insertions, and V_{cdr3} and J_{cdr3} correspond to the sets of suffixes of V genes and prefixes of J genes that form parts of the CDR3 sequences. All random variables in the model are drawn according to a joint distribution on all the variables.

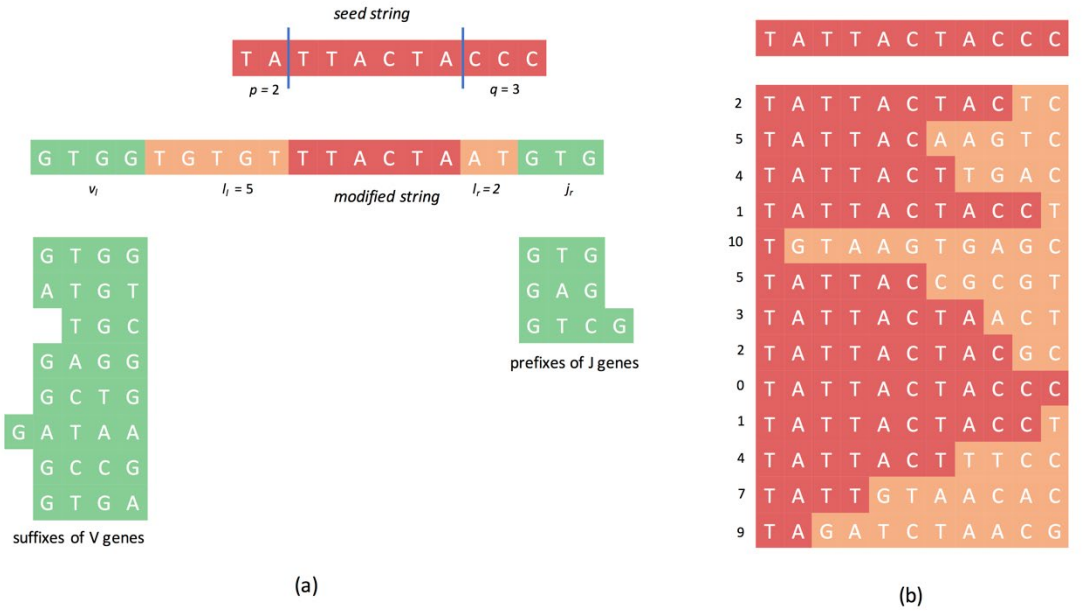

Figure 2.1. Transformation of a seed string representing a D gene into a modified string representing a CDR3 (a) and a set of modified strings generated according to a simple probabilistic model (b). (a) The symbols in red, yellow, and green in the modified string denote the symbols from the truncated seed string, random insertions, and V suffixes/J prefixes, respectively. The sets of V suffixes and J prefixes are shown below the modified string. Note that the sequences shown here are only for illustration and do not correspond to any real genes. (b) In a simple probabilistic model, suffixes of length k are trimmed from the seed string, and the trimmed string is extended by k random symbols, where k (shown by numbers on the left) is chosen uniformly at random. Note that in most cases, there are multiple ways a modified string can be generated from the original string. For example, the first modified string can be generated from the original string by trimming the suffix "CC" and adding the string "TC" or by trimming the suffix "CCC" and adding the string "CTC."

2.2.2 D gene inference and the trace reconstruction problem.

Given a set $C=\{c_1, c_2, c_3,.., c_N\}$ of independently generated instances of the modified strings generated from an unknown set of seed strings $S=\{s_1, s_2,.., s_M\}$, the D genes inference problem is to reconstruct the set S of seed strings. This problem can be thought of as a version of the *trace reconstruction problem* in information theory [31] *i.e.*, reconstruction of an unknown string s given a collection of its *traces* generated according to a given probabilistic model. In the trace reconstruction problem, an unknown string s yields a collection of traces, each trace independently obtained from s by deleting each symbol with a given probability. In the D genes inference problems, traces are generated according to a more complex probabilistic model with multiple parameters.

2.2.3 A simple probabilistic model.

Although the described variant of the trace reconstruction problem represents an adequate probabilistic model for the VDJ recombination, estimating a joint distribution on the variables that accurately mimics the real recombination events is a difficult task. For the sake of simplicity and to develop an intuition for the MINING-D algorithm, we consider a simpler probabilistic model that is based on a single seed string s (representing a single D gene) rather than a set of strings (representing multiple D genes) that gets trimmed only on one side (Figure 1b).

Let s be a seed string in an alphabet $\mathcal{A}$. The seed string generates a modified string c according to the following probabilistic process:

1. A *trimming integer* k is sampled uniformly at random from $[0, |s|]$, and the suffix of s of length k is trimmed.
2. The resulting string is extended by k symbols on the right where each symbol is uniformly selected at random from the alphabet $\mathcal{A}$.

Note that a seed string may generate the same modified string for different values of the trimming integer k. For example, a seed string ATGA may generate a modified string ATCC for k=2 (with probability 1/5*1/16 in the case of 4-letter alphabet $\mathcal{A}$), or a modified string ATCC for k=3

(with probability 1/5*1/64), or a modified string ATCC for k=4 (with probability 1/5*1/256). The probability $P\ (c|s)$ that a seed string s generates a modified string c depends only on the length m of their longest shared prefix and is given by

$$P\ (c|s) = \frac{1}{|s|+1}\sum_{k=0}^{m}\frac{1}{|\mathcal{A}|^{|s|-k}} \tag{1}$$

$$= \frac{1}{(|s|+1)|\mathcal{A}|^{|s|}}\ \sum_{k=0}^{m}|\mathcal{A}|^{k} \tag{2}$$

$$= \frac{1}{(|s|+1)|\mathcal{A}|^{|s|}} \times \frac{|\mathcal{A}|^{m+1}-1}{|\mathcal{A}|-1} \tag{3}$$

$$= K(|s|,|\mathcal{A}|)\ \times\ (|\mathcal{A}|^{m+1}-1) \tag{4}$$

where $K(q,|\mathcal{A}|)$ is a constant given length of the seed string and the alphabet size. Given a set C = $\{c_1, c_2, c_3, \ldots, c_N\}$ of N modified strings independently generated from the same seed string s, the probability that s generates C is computed as

$$P(C|s) = \prod_{i=1}^{N} P(c_i|s) \tag{5}$$

2.2.4 String Reconstruction Problem.

Given a set of modified strings C generated by an unknown seed string, find a string s maximizing $P(C|s)$.

Maximizing $P(C|s)$ is equivalent to maximizing $\prod_{i=1}^{N} K(|s|,|\mathcal{A}|)\ \times\ (|\mathcal{A}|^{m_i+1}-1)$, where m_i stands for the length of the longest shared prefix of s and c_i. Since $K(|s|,|\mathcal{A}|)$ is a constant, it is equivalent to finding a string s that maximizes:

$$\text{score}(C|s) = \sum_{i=1}^{N}\log(|\mathcal{A}|^{m_i+1}-1) \tag{6}$$

Interestingly, if one ignores the "-1" term above, this problem is equivalent to finding a string s that maximizes $\sum_{i=1}^{N} m_i$, the number of "red" cells in the matrix shown in Figure 2.1(b). Given a string s, $\text{score}(C|s)$ can be computed in O($|s|$*N) time.

2.2.5 Greedy algorithm for D gene inference

Although the objective function in (6) can be efficiently maximized (see Supplemental Note: Exact Algorithm for solving the String Reconstruction Problem), it is unclear how to generalize that algorithm for the more complex model with multiple D genes and varying lengths of modified strings. We thus describe a suboptimal greedy algorithm that is easier to extend to cases where the assumptions of the simpler model do not hold. The algorithm starts with an empty string and at step j extends it on the right by the most abundant symbol in C at position j and discards from C the strings that have symbols that are not the most abundant symbols at position j (more details in Supplemental Note: Greedy Algorithm). This procedure is repeated until the length of the resulting string is equal to the length of the seed string s.

To account for the complexities of the VDJ recombination process, we need to modify the greedy algorithm described above. Therefore, for the original D gene inference problem from CDR3 sequences, we describe a heuristic algorithm MINING-D (**M**ethod for **IN**ference of **I**mmu**N**oglobulin **G**enes - **D**) inspired by the above greedy algorithm and considering the complexities of real CDR3s.

2.2.6 MINING-D algorithm

Figure 2.2 presents the outline of the MINING-D algorithm. Although D genes typically get truncated on both sides during the VDJ recombination process, their *truncated substrings* are often present in the newly recombined genes, and, hence, the CDR3s. Therefore, the truncated substrings of D genes are expected to be highly abundant in a CDR3 dataset (Figure 2.2). MINING-D first finds highly abundant k-mers in a CDR3 dataset and then iteratively extends them on both sides to recover the entire D gene based on the elevated relative abundances of the extended substrings. We illustrate

the steps of the MINING-D algorithm on a CDR3 dataset constructed from the ERR1759678 sample (MOUSE dataset). The MOUSE dataset corresponds to a pet shop mouse (the strain is unknown) and consists of 124,121 distinct CDR3s. In general, MINING-D does not rely on the productivity of the input sequences and can work on any fragments of the VDJ region (both in-frame and out-of-frame) that cover the entire D gene (as well as short segments of V and J genes).

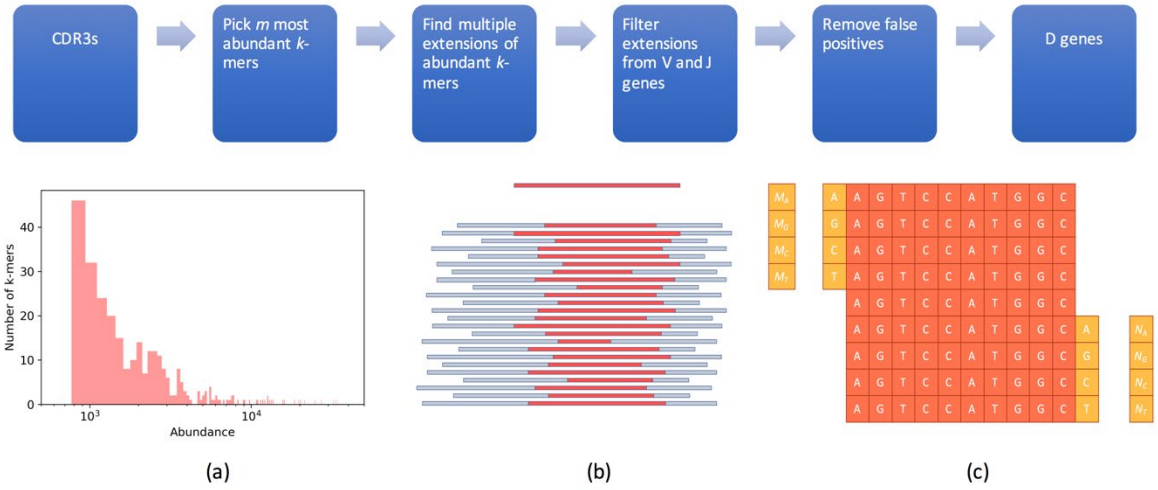

Figure 2.2. Outline of the MINING-D algorithm. (Top) MINING-D pipeline. (Bottom, a) Abundances of the 300 most abundant 10-mers in the MOUSE dataset vary from 770 to 34,451. (Bottom, b) A D gene (top) and its truncated substrings in various CDR3s of varying lengths (bottom). The red part of a CDR3 is the "surviving" substring of the D gene shown at the top whereas the blue part represents the non-D gene part (parts of V and J genes and random insertions). Some substrings of the original D gene, mostly central, are highly abundant. (Bottom, c) A *k*-mer is extended based on the relative abundances of the four shown (*k*+1)-mers on the left and the four shown (*k*+1)-mers on the right.

2.2.7 Seed selection.

MINING-D starts with the *m* most abundant *k*-mers in the CDR3 dataset referred to as *seeds* (default $k = 10$, the selection of the default value of *m* depends upon the species and is described in Supplemental Note: MINING-D Parameters). Most seeds represent substrings of D genes, or strings that have suffixes of V genes or prefixes of J genes as substrings. Our goal is to extend seeds originating from D genes into full-length D genes and filter out seeds originating from (potentially unknown) V and J genes. The abundances of the $m = 300$ (default value for mice datasets) most abundant 10-mers in the MOUSE dataset ranged from 770 to 34,451 (Figure 2.2).

2.2.8 Extending seed *k*-mers and the stopping rule.

Given a string of length *l*, MINING-D analyzes all its possible extensions on the left and right by a single nucleotide. We test a hypothesis that this string represents the first (last) *l*-mer in some D genes, and thus any nucleotide present immediately on the left (right) of this *l*-mer in CDR3 sequences is a random insertion. If the (*l*+1)-mer resulting by adding the corresponding nucleotide on the left (right) is also a substring of the same D gene, the hypothesis will most likely be rejected, and an extension is made using the most abundant extension symbol.

We start the above procedure with seed *k*-mers. For a highly abundant seed *k*-mer, let the abundances of the four possible extension (*k*+1)-mers on the right be N_A, N_G, N_C, and N_T (Figure 2.2). We assume a probabilistic model in which a random nucleotide is added to the last *k*-mer according to some distribution. The statistic *S*, where

$$S = \sum_{i \in \{A,G,C,T\}} \frac{(N_i - E_i)^2}{E_i}$$

and E_i is the expected abundance under the distribution of the (*k*+1)-mer with the nucleotide *i* added to the right of the *k*-mer is approximately Chi-square distributed with 3 degrees of freedom. We test the null hypothesis that the random nucleotide was added according to a uniform distribution, and, thus, the expected abundances are equal under the null hypothesis. The null hypothesis is accepted or rejected based on the *p*-value of the test. The robustness of the choice of equal abundances of the four (*k*+1)-mers under the null hypothesis can be, to some extent, controlled by choosing a significance threshold to which the *p*-value is compared to accept or reject the hypothesis. Having a low significance threshold will lead to rejection of the hypothesis only when the observed distribution of the abundances of the four (*k*+1)-mers is very different from the uniform distribution, most likely in the case where one of the (*k*+1)-mers is much more abundant than the others (see also Supplemental Note: MINING-D Parameters). The statistical test is run on both the distributions – one with the abundances of the four (*k*+1)-mers corresponding to the extensions on the left and the

other corresponding to the extensions on the right. If one of the two hypotheses is rejected, the k-mer is extended to the most abundant (k+1)-mer corresponding to the rejected hypothesis. If both the hypotheses are rejected, the extension is made corresponding to the hypothesis with a lower p-value of the test. In any case, if the k-mer is extended to a (k+1)-mer, the procedure is repeated until both the hypotheses are accepted. Thus, for every highly abundant seed k-mer, we generate a string containing this k-mer.

2.2.9 Finding multiple extensions of seed k-mers.

Some highly abundant k-mers can be substrings of multiple D genes as shown in Figure 2.3(a). Following the procedure above, if we start with a k-mer that is a substring of multiple D genes, its extension will most likely correspond to the more abundant D gene in the CDR3 dataset (among D genes containing this k-mer). Therefore, sometimes multiple extensions are desired from a single abundant k-mer. However, since it is not clear how to avoid false positives in the case of multiple extensions, the IgScout algorithm [28] uses long seed k-mers (that are unique among all D genes), thus bypassing the multiple extension problem. Although this approach works for species with partially known germline genes, it is unclear how to select k for species with unknown germline genes and short germline genes.

To address this limitation of IgScout, we modified the extension procedure described above. After rejecting a hypothesis at any step (say j for $j \geq k$) and extending the j-mer to the most abundant (j+1)-mer, we further test the hypothesis that the remaining three (j+1)-mers follow a random uniform distribution. If the j-mer was a part of two D genes and the selected (j+1)-mer corresponds to the more abundant D gene among those, the abundance of the (j+1)-mer corresponding to the lesser abundant D gene will still be greater than the (j+1)-mers corresponding to the random insertions. Hence, the hypothesis will be rejected, and in the next step, extensions of both the (j+1)-mers are looked for in an independent manner, leading to multiple extensions from a single abundant k-mer. On the MOUSE dataset, the 300 most abundant 10-mers lead to 544 extensions.

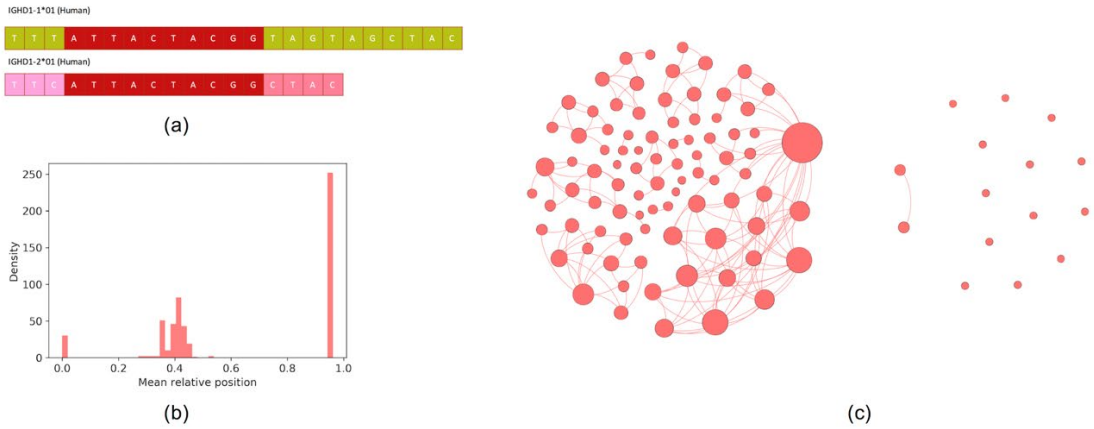

Figure 2.3. Details of MINING-D algorithm. (a) The 10-mer ATTACTACGG is present in two human D gene segments. (b) The mean relative positions of the extensions in the MOUSE dataset. The relative positions of the extensions form three clusters each corresponding to one of the V, D, and J gene segments. (c) Similarity graph on all extensions corresponding to D genes before filtering extensions or clique merging for the MOUSE dataset (left) and after filtering extensions and merging cliques (right). The size of a node represents its degree.

2.2.10 Filtering extensions originating from V and J genes.

Since the CDR3 sequences contain some suffixes of V genes and prefixes of J genes, many highly abundant *k*-mers in the CDR3 dataset originate from these suffixes/prefixes rather than D genes. Therefore, it is important to classify the extensions as corresponding to V, D, or J genes while trying to infer D genes from CDR3 sequences. This problem becomes challenging when the V and J genes are unknown.

Since parts of the V, D, and J genes appear in order in each CDR3 sequence, we use the mean relative position of an extension in the CDR3 dataset to classify it as corresponding to one of the V, D, or J gene segments. We define the relative position of a substring s in a CDR3 sequence c as follows:

$$RP_c(s) = \frac{I_c(s)}{|c|-|s|+1},$$

where $I_c(s)$ is the index of the substring s in the list of all the substrings of length $|s|$ in c ordered from first to last. The normalization by the total number of substrings of length $|s|$ in c is done to compare the relative positions among CDR3s of varied lengths. The relative position of an extension in the entire CDR3 dataset is taken as the mean of the relative positions of the extension in all the CDR3

sequences of which it is a substring. Looking at the relative positions of the extensions of *k*-mers has some advantages over looking at the relative positions of the *k*-mers as explained in the Supplemental Note: Defining Relative Positions. The mean relative positions of the extensions of abundant 10-mers from the MOUSE dataset are shown in Figure 2.3(b). Since the central cluster most likely corresponds to the extensions corresponding to the D genes, MINING-D discards the extensions in the left and right clusters.

However, not all the unique extensions in the central cluster correspond to different D genes. The extensions are first filtered according to the method described in the Supplemental Note: Removing Unidirectional Extensions. Out of the 544 extensions corresponding to the MOUSE dataset, 123 remained after filtering out *unidirectional* extensions, out of which only 52 were unique. Of these 52, only 19 were in the central cluster.

2.2.11 Removing false positives.

To reduce the number of reconstructions per D gene, we construct an undirected *similarity graph* on the inferred extensions. Two extensions are adjacent in the graph if they are *similar*. The distance between extensions e_1 and e_2 is defined as $Dist(e_1,e_2) = \min(|e_1|, |e_2|) - |substring(e_1,e_2)|$, where $substring(e_1,e_2)$ is the longest common substring of e_1 and e_2. It denotes the number of nucleotides, at the edges, of one extension that need to be changed or deleted to transform it to the other extension or a substring of the extension. The larger this number, the more dissimilar the extensions are. We connect extensions e_1 and e_2 with an edge if $Dist(e_1,e_2)$ does not exceed a threshold *maxDist* (the default value is 2).

Cliques in the constructed graph correspond to groups of highly similar extensions. For every clique in the graph, we find the longest common substring among the extensions and extend it to form a new string. This new string then replaces all the extensions that formed the clique. After this clique merging procedure, only 15 of the 19 extensions remained in the MOUSE dataset. Figure 2.3(c) shows the similarity graph among the extensions before and after filtering unidirectional

extensions and merging cliques for the MOUSE dataset. To generate a comprehensive database of D genes from multiple datasets corresponding to different individuals of the same species and health condition, inferred D genes from the datasets were put together and processed (substrings were removed and similar D genes were merged).

2.2.12 Computing usage of the inferred D genes.

Given a set of D genes, we say that a *k*-mer is *unique* if it occurs in a single D gene from this set. We limit attention to *k*-mers that are at least K_{min} nucleotides long (default value $K_{min} = 8$) and say that a CDR3 sequence *c* is *formed* by a D gene *d* if c contains a unique *k*-mer from *d* but does not contain unique *t*-mers from other D-genes for $t \geq k$. A CDR3 sequence is *traceable* if it is formed by a D gene and *non-traceable* otherwise. The usage of a D gene is defined as the proportion of the traceable CDR3 sequences that were *formed* by the D gene.

2.3 Results

2.3.1 Immunosequencing datasets.

We analyzed 588 immunosequencing datasets from 14 publicly available NCBI projects:

- **Human**
    1. **Allergy**. 24 peripheral blood mononuclear cell (PBMC) and bone marrow datasets from six allergy patients from the NCBI project PRJEB18926 [32].
    2. **Flu vaccination**.
        a. 95 datasets taken at different times after vaccination from the NCBI project PRJNA324093 corresponding to different types of cells from eight individuals [33].
        b. 18 PBMC datasets taken either before vaccination or at least two weeks after the vaccination from three individuals from the NCBI project PRJNA349143 [34].

3. **Healthy**. 28 PBMC datasets corresponding to either IgG or IgM isotypes from three individuals from the NCBI project PRJNA430091 [35].
4. **Cord Blood**. 6 datasets corresponding to cord blood samples from five individuals from the NCBI project PRJNA393446.
5. **Intestinal**. 35 datasets from seven individuals corresponding to different types of isotypes and cell types from the tissues ileum mucosa and colon mucosa from the NCBI project PRJNA355402 [36].
6. **Multiple Sclerosis**. 32 datasets from four multiple sclerosis patients corresponding to various tissues with different stages of the disease from the NCBI project PRJNA248475 [37].
7. **Hepatitis B**.
    a. 142 datasets corresponding to IgG isotype and various cell types from nine individuals following a Hepatitis B primary vaccination from the NCBI project PRJNA308566.
    b. 107 datasets corresponding to IgG and IgM isotypes and various cell types from nine individuals following a Hepatitis B booster vaccination from the NCBI project PRJNA308641.

- **Mouse**. 71 datasets from various cell types (pre-B cells, naive B cells, plasma cells) of 20 untreated and antigen-immunized mice from the strain C57BL/6J, and naive cells of four Balb/c mice and three pet mice from the NCBI project PRJEB18631 [38].
- **Macaque**. 7 datasets from three Indian and four Chinese origin rhesus macaques from the NCBI project PRJEB15295 [24].
- **Camel**. 6 datasets corresponding to the VH and VHH isotypes from three camels from the NCBI project PRJNA321369 [39].
- **Rat**. 10 datasets, each corresponding to an immunized rat of Wistar strain from the NCBI project PRJNA386462 [40].

- **Rabbit**. 7 datasets corresponding to spleen and PBMC of three New Zealand rabbits at different stages of a multi-step immunization from the NCBI project PRJNA355270 [41].

Some immunosequencing datasets in a project represent different samples of immunosequencing data from the same environment representing the same individual, tissue, isotype, etc. (e.g., Donor 1, bone marrow sample 1 and Donor 1, bone marrow sample 2). We merged sequences in such datasets to construct a larger CDR3 dataset corresponding to the same environment. Supplemental Note: Immunosequencing Datasets presents summaries of all immunosequencing datasets analyzed in this study. Meta-categories of these datasets were created for different types of analyses and are shown in Table 2.1.

Table 2.1 Meta-categories of datasets.

Meta-category	Datasets	Condition(s)
Healthy PBMC	Allergy	PBMC Either before vaccination or at least 2 weeks after (flu vaccination)
	Flu Vaccination	
	Healthy	
Healthy PBMC & Bone Marrow (BM)	Allergy	PBMC or Bone Marrow Either before vaccination or at least 2 weeks after (flu vaccination)
	Flu Vaccination	
	Healthy	
Tissue Specific	Intestinal	All
	Cord Blood	
Stimulated Datasets	Flu Vaccination	All
	Hepatitis B	
	Multiple Sclerosis	
Non-human	Mouse	All
	Macaque	
	Camel	
	Rat	
	Rabbit	

2.3.2 Constructing CDR3 datasets.

For each immunosequencing dataset, we computed CDR3s using the DiversityAnalyzer tool [42]. Since DiversityAnalyzer uses the set of known V and J genes to compute CDR3s and since V and J genes for camel and macaque are unknown, we used human V and J genes to construct CDR3s for these species. Since some CDR3s may be affected by sample preparation errors, we grouped

CDR3s differing by at most 3 mismatches and constructed a consensus CDR3 for each group as described in [28]. Constructing consensus CDR3s also helps concentrate on only the recombinant diversity (and not SHMs) of immunosequencing datasets by removing CDR3s with SHMs to some extent. We ignored datasets with less than 15,000 consensus CDR3s for the inference of D genes.

2.3.3 Known D genes.

The ImMunoGeneTics (IMGT) database [15] contains information about human, mouse, rat, and rabbit germline D genes. We used the IMGT D genes of crab-eating macaques for rhesus macaque analysis and the IMGT D genes of alpacas for camel analysis. Table 2 provides information about the D genes of all these species.

Table 2.2. Information about the D genes in the IMGT database for various species.

Species	# D genes (allelic variations)	# distinct sequences	range of lengths of D genes
Human	27 (7)	32	11 - 37
Mouse	31 (8)	28	10 - 29
Rat	35 (2)	35	10 - 29
Rabbit	14 (0)	10	24 - 42
Crab-eating macaque	40 (0)	35	11 - 42
Alpaca	8 (0)	8	11 - 34

2.3.4 Inferred D genes.

For inference of human D genes, PBMC datasets from Healthy, PBMC Flu Vaccination datasets taken either before vaccination or at least two weeks after vaccination, and PBMC datasets from Allergy datasets were considered (Healthy Human PBMC datasets, Table 2.1). This was done so as to not include any disease specific changes in the repertoire for inference of D genes. For all other species, all available datasets were used. All inferred genes from an immunosequencing dataset (or multiple datasets) were classified into the following categories based on the IMGT database:

- Inferred genes in IMGT – the inferred gene is either (i) the same as a known D gene or a known variation, or (ii) a substring of a known D gene or a known variation, or (iii) a substring of a known D gene or a known variation extended by at most *extension* extra nucleotides at the start and/or the end of that substring (the default *extension* = 3).
- Novel variation – the inferred gene differs from a known D gene in the database with percent identity > 75%.
- Novel gene – the inferred gene has percent identity < 75% compared to all known D genes in the database.

Table 2.3 presents information about the number of inferred D genes from each species and their classification into one of the categories above. To benchmark the performance against IgScout, we compared the results of MINING-D and IgScout on all Allergy datasets from the project PRJEB18926 and many non-human datasets (see Supplemental Note: Benchmarking MINING-D against IgScout). For human datasets, IgScout failed to reconstruct seven D genes from the IMGT database from all the datasets, whereas MINING-D only missed three genes.

Table 2.3. Information about inferred D genes. The number of novel genes and variations validated using genomic data (procedure described later) are shown.

Species #Individuals	IMGT Database	# Inferred genes	# Inferred genes in IMGT	# Novel variations (validated)	# Novel genes (validated)
Healthy Humans 20	Human	38	25	8 (2)	5 (0)
Untreated + Immunized Mice 27	Mouse	24	18	5 (1)	1 (0)
Immunized Wistar Rats 1	Rat	16	12	4 (3)	-
Rhesus Macaques 7	Crab Eating Macaque	25	17	6 (6)	2 (2)
Bactrian Camels 3	Alpaca	13	1	12 (8)	-
Rabbit 3	Rabbit	18	3	13 (3)	2 (0)

2.3.5 Novel Variations.

Among the 38 (m = 600) inferred D genes from the Healthy Human PBMC datasets corresponding to 20 individuals, 8 were labeled as novel variations, including four variations of the gene IGHD3-10*01, two variations of the gene IGHD3-22*01, and single variations of the genes IGHD2-2*01 and IGHD3-16*02. Table 2.4 presents the sequences of the validated (validation procedure described later) novel variations of genes in Human and other datasets. Note that although only the sequence TTATGATTAC**A**TTTGGGGGAGTTATCGTTAT was inferred as a novel variation of the gene IGHD3-16*02 (N_Var (IGHD3-16*02)-0) from immunosequencing data, the full sequence **GTA**TTATGATTAC**A**TTTGGGGGAGTTATCGTTAT**ACC** was found in genomic reads (more details in next subsection). Information about all inferred variations (including variations that could not be validated using genomic data) is presented in Supplemental Note: Novel Variations.

For rhesus macaques, the two novel genes inferred seem to be two variations of the same novel gene with the following sequences:

N_Gene-0 TACAAT**T**TTTGGA**G**TGGTTAT

N_Gene-1 ATTACAAT**A**TTTGGA**C**TGGTTATTAT

The sequences of these genes found in the genomic data from different individuals of the same species are as follows:

N_Gene-0 GTATTACAAT**T**TTTGGA**G**TGGTTATTA**C**ACC

N_Gene-1 GTATTACAAT**A**TTTGGA**C**TGGTTATTA**T**ACC.

Table 2.4. Novel variations of D genes validated using genomic data from human, camel, rhesus macaque, mouse, rat, and rabbit datasets. "Original" refers to the sequence in the IMGT database. In three of the inferred sequences, there is an extra nucleotide at the end that was not found in the genomic reads, e.g., the novel variation inferred from mice datasets TTTATTACTACG**A**T**G**GTAGCTACg is only present as TTTATTACTACG**A**T**G**GTAGCTAC in the genomic reads. Other polymorphisms that were found using genomic validation of the inferred genes are underlined. For example, GATACAG**C**GGGTACAGT was inferred by MINING-D as a variant of the macaque gene IGHD5S3*01, but the whole sequence G**G**GGATACAG**C**GGGTACAGTTAC was found in the genomic reads.

Human

IGHD3-10*01

```
Original    GTATTACTATGGTTCGGGGAGTTATTATAAC
N_Var-3     GTATTACTATGGTTCAGGGAGTTATTATAAC
```

IGHD3-16*02

```
Original  GTATTATGATTACGTTTGGGGGAGTTATCGTTATACC
N_Var-0   ---TTATGATTACATTTGGGGGAGTTATCGTTAT---
```

Camel

IGHD3*01 (Alpaca)

```
Original   GTATTACTACTGCTCAGGCTATGGGTGTTATGAC
N_Var-1    ----GACTGCTATTCAGGCTCTTGGTGTTATG--
N_Var-0    ---TGACTACTGTTCAGGCTCTTGGTGT------
```

IGHD2*01 (Alpaca)

```
Original    ACATACTATAGTGGTAGTTACTACTACACC
N_Var-1     --ATATTGTAGTGGTGGTTACTGCTAC---
N_Var-0     GCATACTATAGTGGTGGTTACTAC------
```

IGHD4*01 (Alpaca)

```
Original    TTACTATAGCGACTATGAC
N_Var-1     CTACTATAGCGACTATG--
N_Var-0     CTACTATAACGAATATG--
```

IGHD6*01 (Alpaca)

```
Original    GTACGGTAGTAGCTGGTAC
N_Var-2     GTACGGTGGTAGCTGGTAC
```

IGHD5*01 (Alpaca)

```
Original    AGACTACGGGTTGGGGTAC
N_Var-0     ----TATGGGTT-GGGTAC
```

Rhesus Macaque

IGHD1S39*01

```
Original    GGTATAGTGGGAACTACAAC
N_Var-0     -----AGTGGGAGCTAC---
```

IGHD3S18*01

```
Original    GTACTGGGGTGATTATTATGAC
N_Var-0     --ACTGGAGTGATTATTA----
```

IGHD5S3*01

```
Original    GTGGATACAGTGGGTACAGTTAC
N_Var-0     -G-GATACAGCGGGTACAGT---
```

IGHD2S11*01

```
Original    AGAATATTGTAGTAGTACTTACTGCTCCTCC
N_Var-0     --C--ATTGTAGTGGTACTTACTGCTATG--
```

IGHD2S17*01

```
Original    AGAATACTGTACTGGTAGTGGTTGCTATGCC
N_Var-0     ----TACTGTACTGGTAGTGGTTGCTAC---
```

IGHD3S23*01

```
Original    GTATTACTATGATAGTGGTTATTACACCCACAGCGT
N_Var-0     ---TTACTATGGTAGTGGTTATTAC-----------
```

Mouse

IGHD1-1*01

```
Original    TTTATTACTACGGTAGTAGCTAC-
N_Var-0     TTTATTACTACGATGGTAGCTACg
```

Rat

IGHD1-3*01

```
Original    TTTTTAACTATGGTAGCTAC
N_Var-0     -TTTTAACTACGGTAGCTAC
```

IGHD1-9*01

```
Original    TACATACTATGGGTATAACTAC-
N_Var-1     --CATACTACGGGTATACCTACg
```

IGHD1-12*02

```
Original    TTTATTACTATGATGGTAGTTATTACTAC-
N_Var-0     -TTATTACTATGATGGTACTTATTACTACg
```

Rabbit

IGHD6-1*01

```
Original   --------------GTTACTATAGTTATGGTTATGCTTATGCTACC
N_Var-4    GTTACTATACTTATGGTTATGCTGGTTATGCTTATGCTACC
N_Var-3    GTTA------TGCTGGTTATGCTGGTTATGGTTATGCTACC
```

IGHD1-1*01

```
Original    GCATATACTAGTAGTAGTGGTTATTATATAC
N_Var-2     GCATATGCTAGTAGTAGTGGTTATTAT----
```

2.3.6 Validation of novel D gene variations using Whole Genome Sequencing data.

To validate novel genes and variations discovered by MINING-D, we downloaded genomic reads for all analyzed species and searched for the occurrences of the novel genes and variations in these reads (see details in Supplemental Note: Finding D genes in Whole Genome Sequencing Data). Since paired genomic and immunosequencing datasets were not available, genomic and immunosequencing reads came from different individuals (Table 2.5). We consider an inferred novel D gene or variation validated if it is present in at least 2 reads and is surrounded by RSS motifs on both sides. Table 2.5 provides details of the downloaded data and information about validated variations.

Table 2.5. Genomic data used for validating discovered D gene variations. The last column describes the number of datasets in which the novel sequences were found in genomic reads and the range of number of reads in which the sequences were found. For rhesus macaques, we chose only 4 datasets out of the 1318 in the NCBI project PRJNA382404 for analysis.

Species	Project	Description	Datasets	Novel variations/genes found in genomic reads	# datasets (# reads)
Human	PRJNA427604	WES of PBMC (ESCC - cohort, China)	40	N_Var (IGHD3-10*01)-3 N_Var (IGHD3-16*02)-0	5 (8-14) 6 (30-58)
Mice	PRJEB18467	WGS of mus musculus	32	N_Var (IGHD1-1*01)-0	19 (1-10)
Bactrian Camel	PRJNA276064	WGS of Old world camels	7	N_Var(IGHD2*01)-0 N_Var(IGHD2*01)-1 N_Var(IGHD3*01)-0 N_Var(IGHD3*01)-1 N_Var(IGHD4*01)-0 N_Var(IGHD4*01)-1 N_Var(IGHD5*01)-0 N_Var(IGHD6*01)-2	2 (2) 6 (4-17) 2 (2-6) 7 (1-16) 2 (4-5) 6 (7-15) 6 (2-13) 7 (5-21)
Rhesus Macaque	PRJNA382404	WGS of rhesus macaques	4/1318	N_Gene-1 N_Gene-1-0 N_Var (IGHD1S39*01) N_Var (IGHD3S18*01) N_Var (IGHD5S3*01) N_Var (IGHD2S11*01) N_Var (IGHD2S17*01) N_Var (IGHD3S23*01)	4 (9-27) 2 (8-9) 1 (18) 2 (8-21) 4 (13-28) 1 (6) 4 (8-30) 3 (12-24)
Wistar Rats	PRJNA479378	WGS of wistar rats	10	N_Var (IGHD1-12*02)-0 N_Var (IGHD1-3*01)-0 N_Var (IGHD1-9*01)-1	10 (1-9) 10 (2-18) 10 (2-8)
Rabbit	PRJNA242290	WGS of rabbits and hares to survey for domestication signals.	24	N_Var (IGHD1-1*01)-2 N_Var (IGHD6-1*01)-3 N_Var (IGHD6-1*01)-4	23 (1-14) 19 (1-9) 11 (1-6)

2.3.7 Usage of D genes.

We analyzed the usage of all IMGT D genes and validated novel genes/variations in Healthy PBMC datasets. 54.1% of CDR3s on average were traceable. The usage of all genes is mostly consistent across individuals, although there are a few deviations for certain individuals owing to their germline variations (Figure 2.4). Potential deletion polymorphisms involving multiple contiguous IGHD genes, as reported in the past [13, 43], can also be seen in Figure 2.4. Donor 10 likely has a deletion of genes D3-3 – D6-6 and donor 11 likely has a deletion covering genes D3-22, D5-24, and D1-26.

To analyze the relative usage of a variant of a D gene (known or novel) against other variants of the same gene, we also included the bone marrow datasets and plotted the variant usage in Healthy PBMC BM (Table 2.1) datasets (Figure 2.5). We found that extensive SHMs in IgG repertoires may lead to a misclassification of alleles for some genes e.g. IGHD3-16 and IGHD2-8. For these genes, we accurately compute the D gene allele usage using decoy alleles as explained in the next subsection.

In the Stimulated datasets, some differences were seen in the usage of D genes in IgG and IgM datasets (see Supplemental Note: D gene usage). For instance, in the Hepatitis B datasets, 65.4% and 45.9% CDR3s were traceable on average in datasets corresponding to IgM and IgG isotypes, respectively. The usage of some genes differs in datasets corresponding to IgG and IgM isotypes from the same individual (Figure 2.6). For example, genes IGHD1-26*01, IGHD6-13*01, and IGHD3-3*01 appear to be used more in the IgM datasets for most individuals whereas IGHD3-9*01 is used more in the IgG datasets.

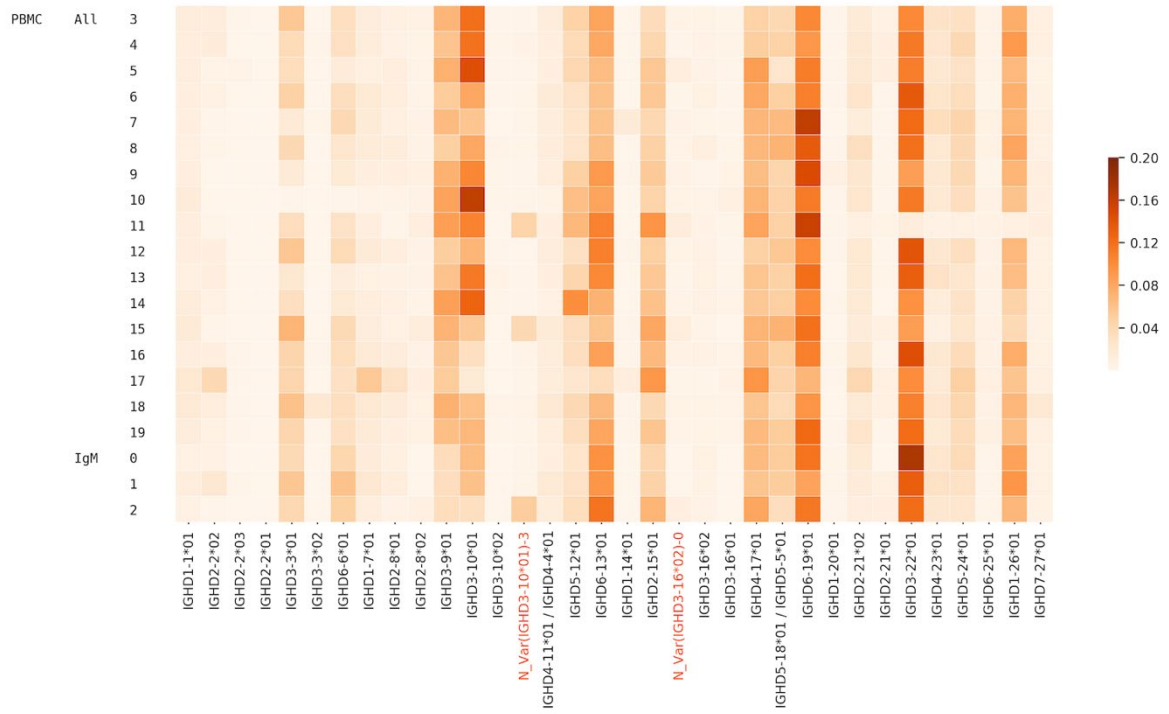

Figure 2.4. Usage of various known and novel genes in various Healthy datasets. Each row corresponds to a different dataset described by the three leftmost columns. The first column denotes the tissue, i.e., PBMC, the second column denotes the isotype (IgM or all/unsorted), and the numbers in the third column represent different individuals. The color in each cell represents the proportion of traceable CDR3s that were formed by a gene on the x-axis in the dataset corresponding to the y-axis. Validated novel variations are highlighted on the x-axis.

2.3.8 Accurate computation of the D gene allele usage.

High SHM rate in IgG datasets can lead to inaccurate labeling of CDR3s in terms of the D gene allele used. Figure 2.5 shows a checkered pattern (particularly in the usage of genes IGHD2-8 and IGHD3-16 in IgM and IgG datasets) for individuals 0 and 1 - while IgM datasets show homozygous state of IGHD3-16 formed by allele 2, IgG datasets show heterozygous state of IGHD3-16 formed by the known allele 1 and the novel allele N_Var-0. This is because the estimated usage of a D gene depends not only on the sequence of that gene but also on the sequences of other genes. If two genes have very similar sequences and only one of them is present in the database, the CDR3 sequences originating from both the genes will get assigned to the one that is present in the database.

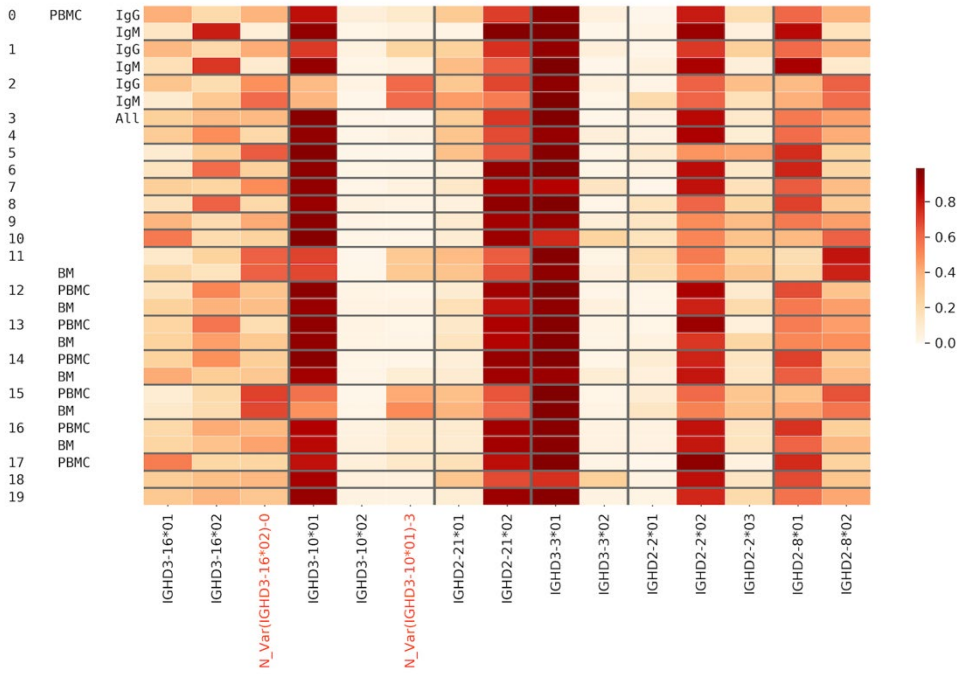

Figure 2.5. Usage of variants of D genes in Healthy PBMC BM datasets. Gray lines separate the plot such that each subplot corresponds to one gene and its variants. Each cell in a subplot represents the proportion of the usage of a variant with respect to the total usage of all variants. Thus, in every subplot, the sum of all rows is 1. The columns on the y-axis tick labels represent the individual, the tissue, and the isotype, respectively.

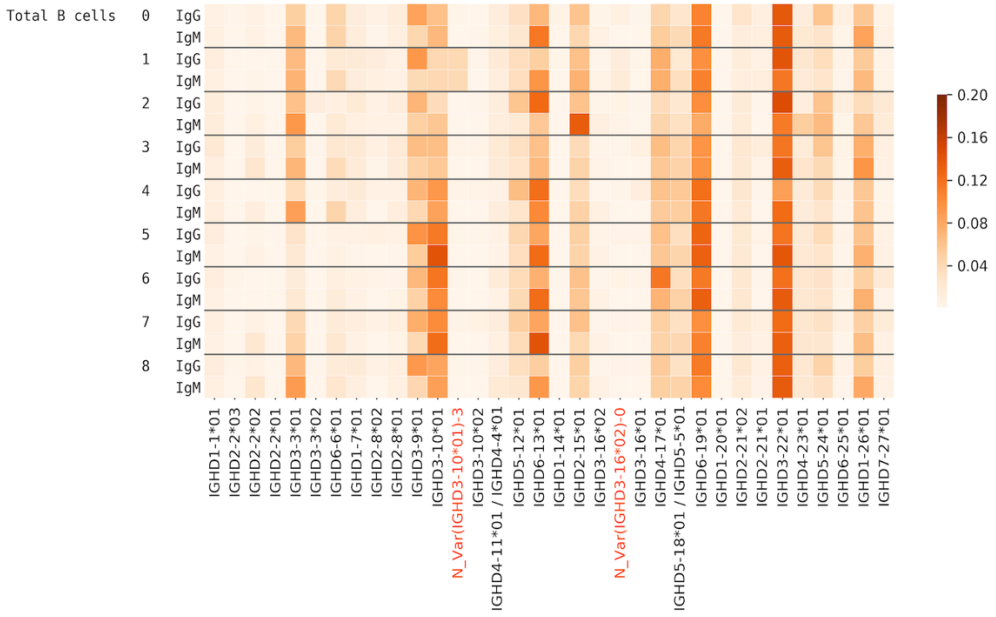

Figure 2.6. Usage of D genes in the Hepatitis B datasets corresponding to the IgG and IgM isotypes from various individuals.

When there are many SHMs, some hypermutated reads (CDR3s) can get assigned to one of the allelic variations if only a few (2-3, usually germline) are present in the database. Since the usage of alleles of a D gene is calculated in terms of proportion of the total usage of the D gene, even a small number of hypermutated CDR3s that got assigned to a wrong allele (because not all possible variations of the gene were in the database) can show up as a considerable proportion of the total usage, particularly if the total usage is small. This is what happened in the cases of genes IGHD3-16 and IGHD2-8 (Figure 2.5).

To circumvent this issue, we added artificial alleles of IGHD2-8, IGHD3-16, and IGHD3-10 to the D genes database to check if all the CDR3s that were assigned to alleles in the IgG datasets for subjects 0, 1, and 2 (Figure 2.5) would still be assigned to the same alleles in the presence of these false variations. We added 61 alleles for IGHD3-16 that are possible with mutations at the highlighted sites in Figure 2.7.

```
IGHD3-16*01   GTATTATGATTACGTTTGGGGGAGTTATGCTTATACC
IGHD3-16*02   GTATTATGATTACGTTTGGGGGAGTTATCGTTATACC
N_Var-0       GTATTATGATTACATTTGGGGGAGTTATCGTTATACC
```

Figure 2.7. Alleles of the gene IGHD3-16.

The results of D gene labeling are shown in Figure 2.8. Most of the CDR3 reads that were falsely assigned were distributed among the false alleles whereas the ones which were correctly assigned did not.

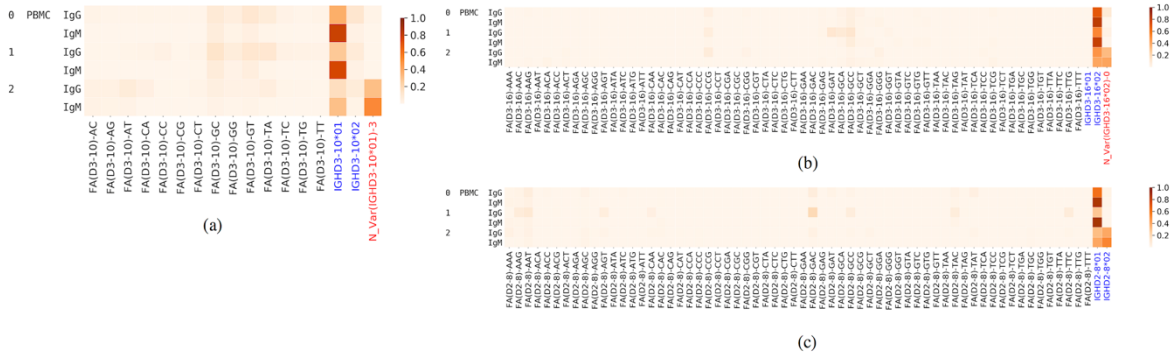

Figure 2.8. Allelic variant usage for genes IGHD3-10 (a), IGHD3-16 (b) and IGHD 2-8 (c). FA stands for false alleles. The alleles listed in IMGT are shown in blue. Novel inferred genes are shown in red.

For genes IGHD3-16 and IGHD2-8, the total usage was much smaller than other genes e.g. IGHD3-10. That is why the CDR3s incorrectly assigned to alleles made up a considerable proportion of the total number of CDR3s that were assigned to all alleles of the gene. Figure 2.8(a) illustrates that results for genes with a slightly higher usage are similar to results in Figure 2.5.

2.3.9 Haplotyping heterozygous V genes using D genes.

To support the inferences of the novel alleles found in the Healthy datasets, we used them for inference of haplotypes of V genes. Haplotype inference, whenever subjects are heterozygous with respect to some genes, can lend support to the identification of novel alleles of the germline genes [44]. We analyzed two Rep-seq datasets corresponding to individual 2 in Figure 2.4 and individual 5 from the Intestinal datasets (see Supplemental Note: D gene usage, Figure A.10). For each individual, we selected V genes that are present in corresponding Rep-seq datasets in the form of at least two allelic variants. To minimize the impact of the sample preparation artifacts and SHMs, we ignored alleles with low usage (<1000 distinct CDR3s). As a result, we selected 12 and 9 heterozygous V genes for individuals 2 and 5, respectively (Table 2.6). Afterward, we extracted distinct CDR3s corresponding to each of the selected alleles and identified D genes in them. The joint usage of V and D gene alleles allows us to identify haplotypes of V genes and 4 heterozygous D genes (including novel alleles of IGHD3-10 and IGHD3-16) in individuals 2 (Figure 2.9) and 5 (Figure 2.10).

We could not infer haplotypes using IGHD2-2 gene because differences between its alleles are concentrated in the start of the gene that is often truncated. We also did not use gene IGHD3-3 that is homozygous in both individuals. In individual 2, we could not infer haplotypes for the 4 out of 12 selected V genes: IGHV1-3, IGHV2-5, IGH3-23, and IGHV3-30. In individual 5, we could not infer haplotypes of IGHV1-69. We assume that it may be caused by the presence of these genes in several copies and SHMs (in individual 2).

Within an individual, haplotypes of the remaining V genes are consistent across all heterozygous D genes. Thus, haplotyping results lend additional support for novel alleles of D genes and prove that heterozygous D genes can be used for haplotyping the IGH locus.

Table 2.6. Abundant heterozygous IGHV genes in individual 2 from Figure 4 and individual 5 from Figure A.10. For individual 2, only the IgM datasets were used. For individual 5, only the naive datasets were used.

V gene	Allele	# distinct CDR3s	
		Individual 2, Figure 4	Individual 5, Figure A.10
IGHV1-2	2	-	4152
	4	-	1299
IGHV1-3	1	7166	-
	2	2175	-
IGHV1-46	1	2154	2717
	3	1058	1445
IGHV1-69	1	-	7872
	6	8820	5194
	9	1914	-
	12	-	3052
IGHV2-5	1	2667	-
	9	2621	-
IGHV3-7	1	2841	-
	3	2276	-
IGHV3-11	1	2569	-
	3	1114	-
IGHV3-15	1	2263	-
	7	2339	-
IGHV3-23	1	1428	-
	4	26138	-
IGHV3-30	1	-	2621
	3	1617	2784
	18	6245	-
IGHV3-48	1	3559	1581
	2	3538	-
	3	-	2753
IGHV3-53	1	2495	-
	2	1048	-
IGHV4-31	1	-	1488
	3	-	8183
IGHV4-39	1	-	6732
	7	-	3448
IGHV4-61	1	-	2828
	2	-	7271
IGHV5-51	1	9561	17635
	3	2143	4082

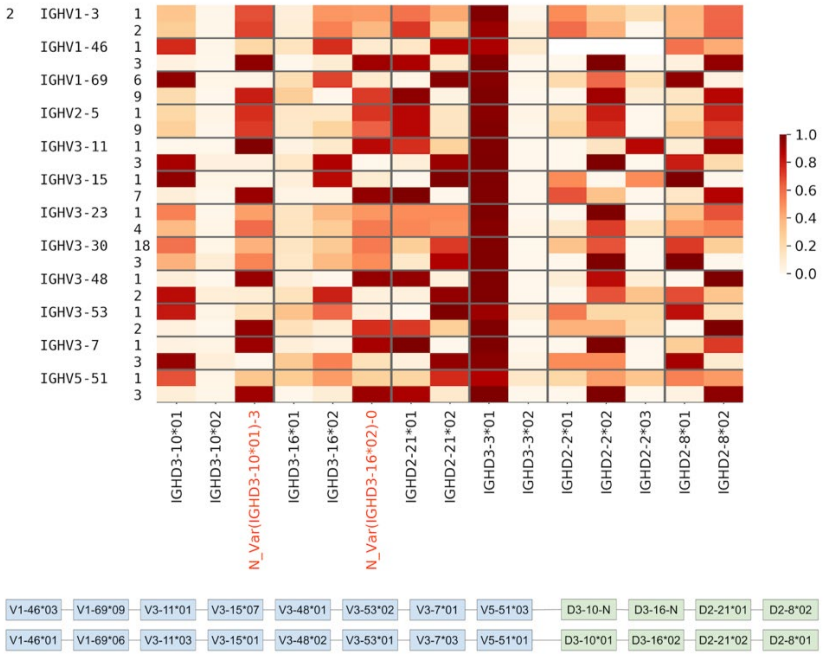

Figure 2.9. Haplotypes of IGHV genes for individual 2 from Figure 4. (Upper) Joint usage of V and D gene alleles**.** Alleles of V genes are shown at the third column on the left. A cell corresponding to allele X of gene V and allele Y of gene D shows the number of distinct CDR3s derived from alleles X, Y normalized by the total number of distinct CDR3s derived from allele X and gene D. Gray lines separate the plot such that each subplot corresponds to one gene and its alleles. (Lower) Haplotypes of IGHV are inferred according to the pairings of alleles of V and D genes supported by the maximum number of CDR3s.

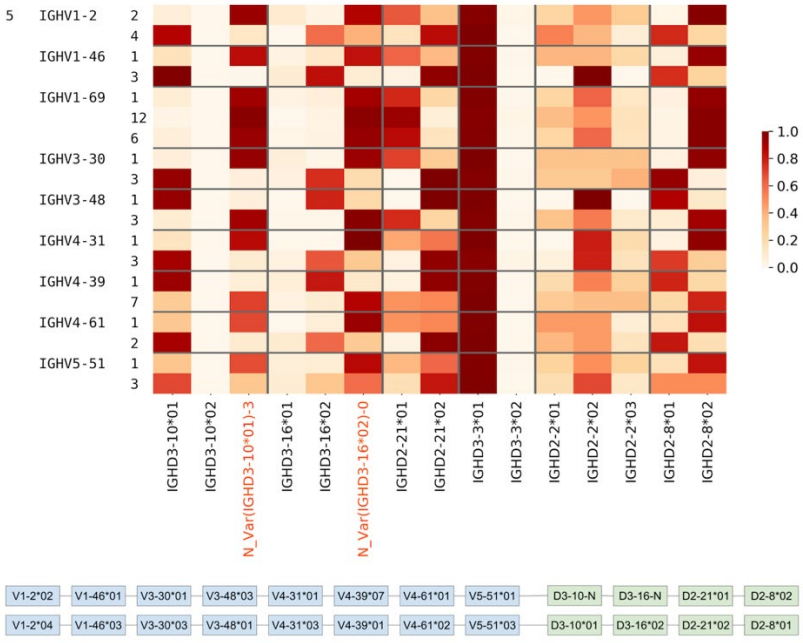

Figure 2.10. Haplotypes of IGHV genes for individual 5 from Figure A.10 (see legend for Figure 2.9).

2.3.10 Overused D genes

To see any potential association between the usage of a D gene and an environment (a health condition, a tissue, or a cell type), we analyzed the usage of D genes in Stimulated and Tissue-specific datasets. We use the gene usage profiles in the Healthy PBMC datasets as a reference and compare the D gene usage profiles in other datasets.

We say that a gene is *overused* in a dataset if the usage of the gene in that dataset is at least twice the maximum usage of that gene in all Healthy PBMC datasets. The ratio of usage of an overused gene to the maximum usage in Healthy PBMC datasets is referred to as *over-usage.* The usages of all IMGT D genes and validated novel variations in Healthy Human PBMC datasets are shown in Figure 2.11. Details on the genes overused in the Flu Vaccination datasets are shown inTable 2.7, and overused genes in other Stimulated and Tissue-specific datasets are shown in Supplemental Note: Over-usage of D Genes. In total, 9 genes were overused in at least 2 datasets of the same type in all Stimulated datasets (Figure 2.12), and 6 genes were overused in at least 2 datasets from the Intestinal datasets (Figure 2.13). These results suggest potential associations between the usage of a D gene and a health condition, tissue, or cell type, although it is difficult to infer statistically significant associations with such a small sample size.

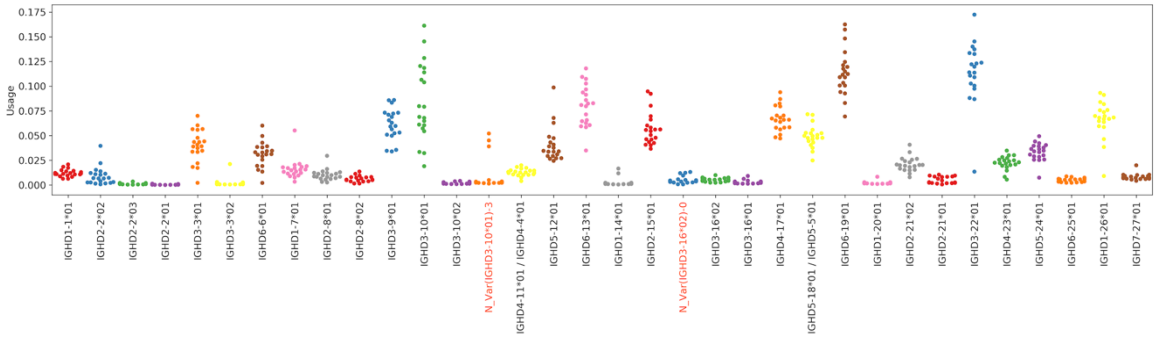

Figure 2.11. D gene usage in all Healthy datasets. Each point above a gene represents a Healthy Human PBMC dataset. To distinguish usages of different genes, adjacent genes are represented by different colors.

Table 2.7. Overused genes in Flu Vaccination datasets. Since the number of datasets in PRJNA324093 is much greater than in other projects, only genes that are overused in at least three different datasets are shown. The over-usage of a gene in a dataset is also shown. For example, the usage of IGHD1-1*01 in HA+ activated B cells for donor 1 is 3.7 times the maximum usage in all Healthy Human datasets.

Gene	Cell type	Donor	Over-usage
IGHD1-1*01	HA+ activated B cell	1	3.7x
		5	13.5x
	HA+ memory cells		5.8x
	HA- activated B cell	7	2.4x
	HA- ASC		4.3x
IGHD2-21*02	HA+ activated B cell	7	8.8x
			6.4x
	HA+ ASC		4.4x
IGHD3-22*01	HA+ activated B cell	4	2.2x
	HA+ ASC		3.4x
	HA- activated B cell		4.4x
	HA+ memory B cell	7	3.2x
IGHD3-9*01	HA+ activated B cell	3	2.0x
	HA- activated B cell	6	2.2x
		7	2.0x
	HA- ASC		2.1x
IGHD4-17*01	HA+ activated B cell	6	9.5x
			8.6x
		7	4.0x
			4.0x
	HA+ ASC	6	9.7x
		7	6.7x
	HA+ memory B cell	6	7.2x

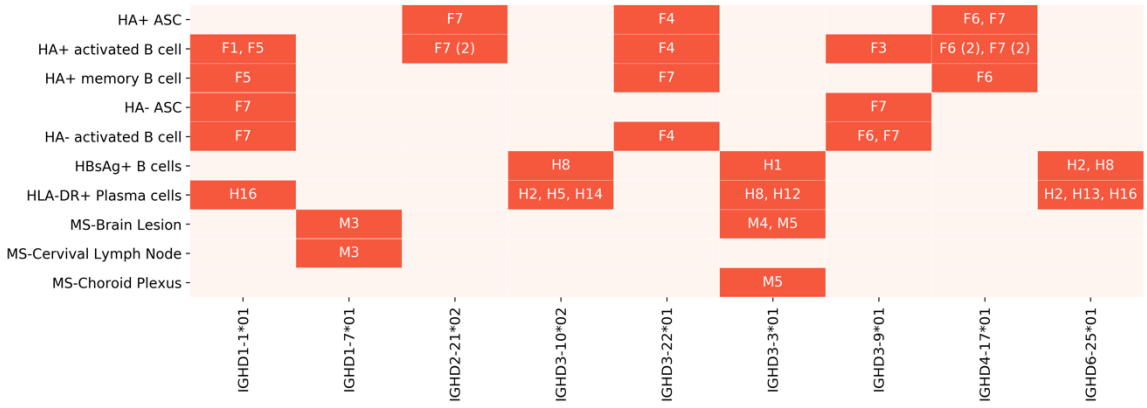

Figure 2.12. Summary of overused genes in Stimulated datasets. The datasets in which each gene is overused are highlighted and annotated with the corresponding individuals. Subjects were prefixed with a letter corresponding to the project – "F" for Flu Vaccination, "M" for Multiple Sclerosis, and "H" for Hepatitis B Vaccination. Some genes were overused in multiple datasets from the same and/or different individuals. The number in parentheses shows the number of datasets from the same individual in which the gene was overused.

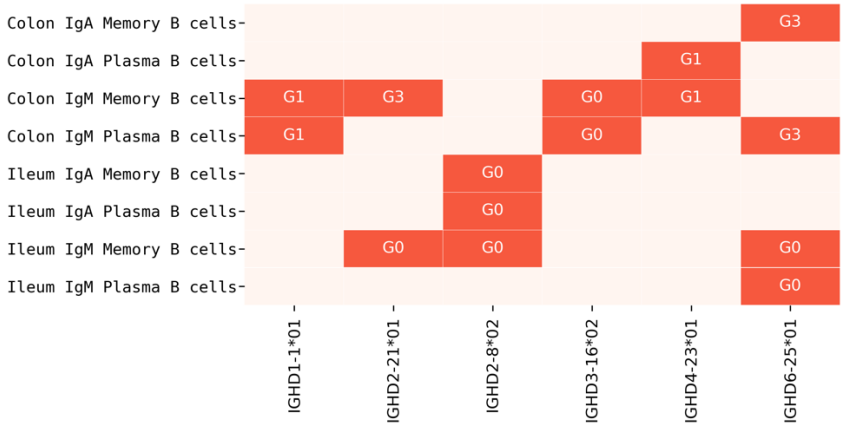

Figure 2.13. Summary of overused genes in Intestinal datasets. The datasets in which each gene is overused are highlighted and annotated with the corresponding individuals. The subjects in the Intestinal Repertoire project were prefixed with "G."

2.3.11 Usage of D genes in the Mouse datasets.

57.4% of CDR3s on average were traceable in each dataset. Figure 2.14 shows the usage of mouse D genes (annotated in IMGT mice and one validated novel variant) in the datasets corresponding to naive B cells of various mice (see also Supplementary Note: D gene usage). The usage of genes among individuals of the same strain is similar. In contrast, the usage of genes among individuals of different strains (Balb/c, C57BL/6J, pet mice) is very different. The gene usages in two of the three pet shop mice (Pet 1 and Pet 2) of unknown strains show a departure from both Balb/c and C57BL/6J strains.

The genes with differential usage in strains Balb/c and C57BL/6J are shown in Figure 2.15. Although the gene IGHD1-1*01 is only listed as a Balb/c gene in the IMGT database, we inferred it in both strains. We inferred genes IGHD1-2*01, IGHD2-10*01/IGHD2-11*01, IGHD2-14*01, and IGHD3-2*01 from only the Balb/c datasets – among these, three of them are listed as Balb/c genes in IMGT whereas IGHD2-14*01 is listed only as a 129/Sv gene. Genes IGHD3-2*02 and IGHD2-5*01/IGHD2-6*01 were inferred from only the C57BL/6J datasets. IGHD3-2*02 is listed as a C57BL/6J gene in the IMGT database. The genes IGHD2-5*01 and IGHD2-6*01 have the same sequence and are listed under the CB.20 strain and C57BL/6J strain, respectively, in the IMGT

database. The results suggest that other than the novel variation that is not listed in the IMGT database for any strain, there are some genes which are listed in the IMGT database of some strains but were also inferred from other strains.

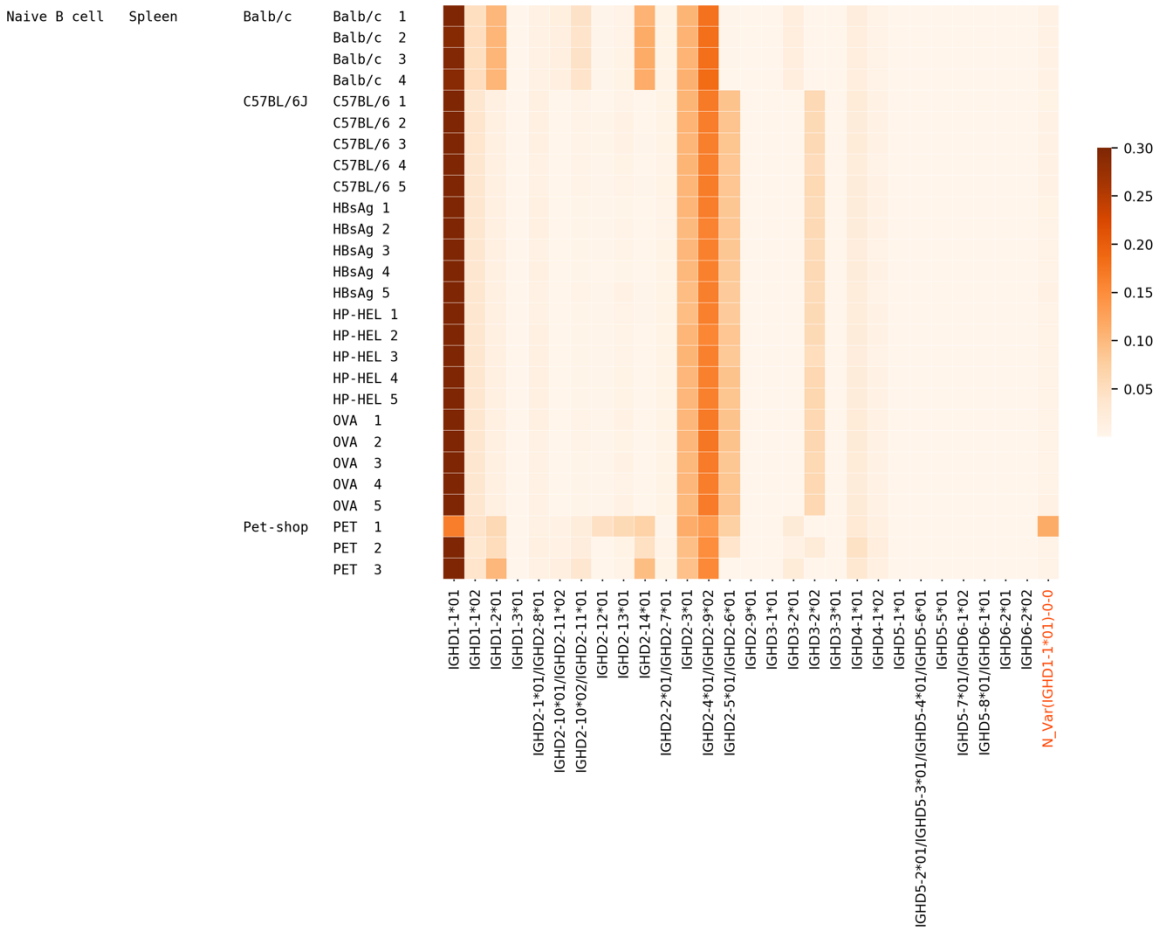

Figure 2.14. Usage of various known and novel genes/variations in MICE datasets. Columns on the left represent cell type, tissue, strain, and individual respectively. OVA, HP-HEL, and HBsAg in the right most column represent the C57BL/6J mice immunized with OVA, HP-HEL, and HBsAg, respectively. For example, OVA 3 represents the C57BL/6J mouse number 3 that was immunized with OVA.

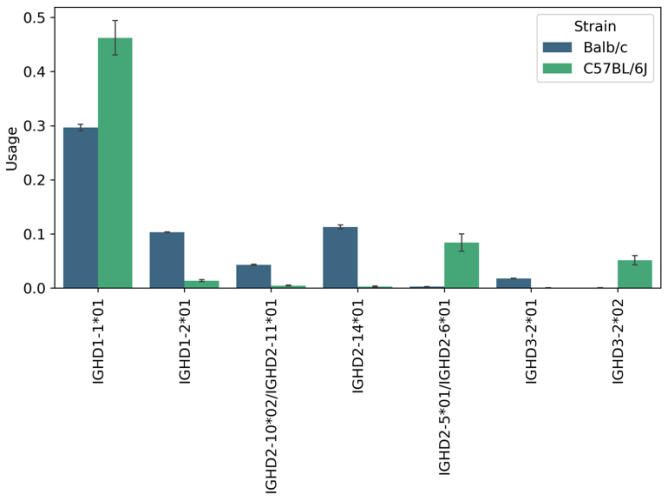

Figure 2.15. Genes with differential usage in Balb/c and C57BL/6J strains. Except IGHD1-1*01, all genes were inferred only in one strain.

2.3.12 Usage of D genes in the camel, macaque, and rat datasets.

31.7%, 52.6%, and 54.3% of CDR3s were traceable on average in the Camel, Macaque, and Rat datasets, respectively (see Supplemental Note: D Gene Usage). The D gene usage profiles were slightly different for the VH and the VHH isotypes within individuals (Figure A.15). For rats, genes belonging to the IGHD2 and IGHD3 families were used much less than in other gene families (Figure A.16). D genes with the highest usage among datasets of a species are shown in Supplemental Note: Highly Used D Genes in Non-human Datasets.

2.4 Discussion

Although inference of *personalized* immunoglobulin V, D, and J genes is now recognized as an important step in the analysis of immunosequencing data [26], inference of D genes presents additional difficulties as compared to inference of V and J genes [5]. Indeed, since D genes undergo exonuclease removals during VDJ recombination (and since they are much shorter than V and J genes), the alignment-based techniques used for V and J gene reconstruction do not work for D gene reconstruction.

Since the most abundant *k*-mers of CDR3s usually originate from D genes, iterative recruitment and extension of abundant *k*-mers in CDR3s (implemented in IgScout [28]) results in *de novo* reconstruction of many germline D genes. The performance of IgScout depends on the value of *k*: selecting a large *k* results in missing short D genes, but selecting a small *k* presents a danger of recruiting *k*-mers that belong to multiple D genes and thus missing some of these genes or producing inaccurate results. For inference of human D genes, IgScout uses $k = 15$ since all 15-mers in known human D genes are unique and all human D genes but one are at least 15 nucleotides long. However, it is unclear how to select the parameter *k* for species with still unknown sets of D genes.

The described MINING-D algorithm does not assume previous knowledge of the lengths of D genes and, unlike IgScout, considers multiple extensions of *k*-mers and thus can use short *k*-mers as seeds (the default value $k = 10$ does not exceed the length of all known D genes). Benchmarking MINING-D on simulated datasets demonstrate high accuracy of the inferred D genes (Supplemental Note: Benchmarking MINING-D on simulated CDR3s).

We applied MINING-D to 588 Rep-seq datasets from various species and inferred 38, 24, 16, 25, 13, and 18 D genes using human, mouse, rat, macaque, camel, and rabbit datasets, respectively. 25 (13), 18 (6), 12 (4), 17 (8), 1 (12), and 3 (15) of human, mouse, rat, macaque, camel, and rabbit D genes were known (novel), respectively. We additionally validated the novel genes and variations using genomic data. Unfortunately, since paired Rep-seq and WGS datasets are currently not available, we could not validate the inferred D genes with genomic data taken from the same individuals. Instead, we downloaded 117 publicly available WGS datasets from different individuals and searched for occurrences of the inferred novel D genes and variations. In total, we validated 25 of the 58 novel D genes/variations. There are multiple reasons why some of the inferred D genes were not validated, e.g., it is difficult to validate a rare allele of a D gene (since paired WGS and Rep-Seq data are not available), inferred gene may be a result of highly abundant SHM rather than a real D gene, etc. We also validated novel alleles of human D genes using haplotyping of heterozygous V genes and showed that haplotypes computed using novel and known D genes are consistent.

Additionally, we benchmarked MINING-D on TCR datasets (Supplemental Note: Benchmarking MINING-D on TCR datasets).

Finally, we analyzed the usage of inferred D genes in diverse Rep-seq datasets and found that it is highly conservative in healthy humans. To see whether a gene is overused in some specific datasets corresponding to a health condition, tissue, and/or cell type, we compared the usage in these datasets against the usage in Healthy Human PBMC datasets as a reference. Based on the results of this comparison, we propose potential associations between some D genes and a health condition, tissue, and/or cell type, albeit the small sample size keeps us from inferring statistically significant associations. In total, we found 9 overused genes among the Flu Vaccination, Multiple Sclerosis, or Hepatitis B Vaccination datasets.

We also analyzed the D gene usage in two mouse strains (Balb/c and C57BL/6J) and demonstrated that the usage of genes among individuals of the same (different) strains was very similar (different). For example, the gene IGHD1-1*01 (which was inferred in both strains) had a much higher usage in the C57BL/6J strain. Since this gene is only listed as a Balb/c gene in the IMGT database, we propose to add it to the database of C57BL/6J genes as well. Similarly, we propose to add IGHD2-14*01 to the Balb/c genes, which is only listed as a 129/Sv gene in the IMGT database.

We demonstrated that high SHM rate may result in erroneous inferences that represent abundant hypermutations rather than novel alleles of D genes (see Supplemental Note: Benchmarking MINING-D on simulated CDR3s). Therefore, inference of novel alleles of D genes (as well as other immunoglobulin genes) must be done from data minimally affected by SHMs (such as naive or IgM / IgD Rep-Seq data) with a follow-up validation of the inferred alleles by genomic data. Using MINING-D, we inferred and validated 25 novel genes/variations in humans, mice, camels, rhesus macaques, rats, and rabbits. We argue that validated novel variations of D genes must be added to standard databases of germline genes to make the analysis of the antibody repertoire data more accurate. In addition, we also analyzed the usage of the known and validated novel D genes in the VDJ recombination process and found that although the gene usage is similar in PBMCs from

healthy individuals, we see some deviations in datasets that are antigen specific. Although, associations between the usage of a D gene and an antigen could not be established due to the low number of samples with a specific data type, our study suggests directions for future research.

2.5 Acknowledgements

Chapter 2 is adapted from Bhardwaj, V., Franceschetti, M., Rao, R., Pevzner, P. A., & Safonova, Y. (2020). Automated analysis of immunosequencing datasets reveals novel immunoglobulin D genes across diverse species. PLOS Computational Biology, 16(4), e1007837. The dissertation author was the primary author of this paper.

Chapter 3

Discovery of fasting molecules using data from non-targeted LCMS

3.1 Introduction

Recent research links fasting to health and longevity [45-47]. Fasting and caloric restriction without malnutrition are linked to numerous benefits in humans including weight loss [48-51], lower blood pressure [48, 49, 52], reduced inflammation [53, 54], and global metabolic improvements in blood lipids and insulin sensitivity [55-59]. Studies in lower organisms suggest that prolongation of lifespan with caloric restriction may be mediated by some carbon-containing metabolites [60], but the identity of these protective molecules remains largely undefined in humans.

Fatty acid esters of hydroxy fatty acids (FAHFAs), a class of recently discovered bioactive lipids [61], favorably modulate diabetes risk and inflammation in adipose tissue in pre-clinical models [61]. Unlike other free fatty acids (FFA) that have been implicated in promoting insulin resistance, ectopic fat deposition, and a pro-inflammatory milieu [62-64], certain species of FAHFAs such as palmitic acid hydroxy stearic acid (PAHSA) appear to exert specific anti-diabetic and anti-inflammatory effects in select animal models of diet-induced obesity (DIO) and insulin resistance

(IR). FAHFAs are present in human and murine foods at very low levels, but their levels can increase in murine AT with fasting [61], suggesting endogenous synthesis. Although the biological effects of 5- and 9-PAHSA have been studied extensively in animal models, data in humans has been relatively limited. There have been no dedicated studies examining the dynamic regulation of FAHFA levels with fasting, refeeding, or specific dietary interventions in humans.

With advancements in accelerated high-performance liquid chromatographic (LC) separation approaches coupled to high-resolution mass spectrometry (MS), routine detection of thousands of unique chemical compounds in a single biosample has become possible. To study the effect of fasting on human health and to see whether there are any metabolites, including the FAHFAs, that increase with fasting and show beneficial effects on human health, we performed a comprehensive systematic measure of potentially protective factors that increase with fasting with the help of nontargeted LC-MS approach. This work was done with the help of, and in collaboration with, Jain Lab at UC San Diego School of Medicine.

3.2 Results

We first examined hourly blood samples taken from 9 obese, non-diabetic individuals during controlled feeding (discovery cohort). About ten thousand metabolites were detected. Metabolite intensities were MAD (median absolute deviation) normalized. To see response to feeding/fasting, we looked at the mean levels of intensities during fasted and fed states for all the detected metabolites. The distribution of the differences between the mean intensities in the fasted and fed states for all the metabolites are plotted in Figure 3.1. We observe that there are some metabolites that increase with fasting, some that decrease with fasting, and others that do not show any response to feeding/fasting. Examples of these categories of metabolites are shown in Figure 3.2.

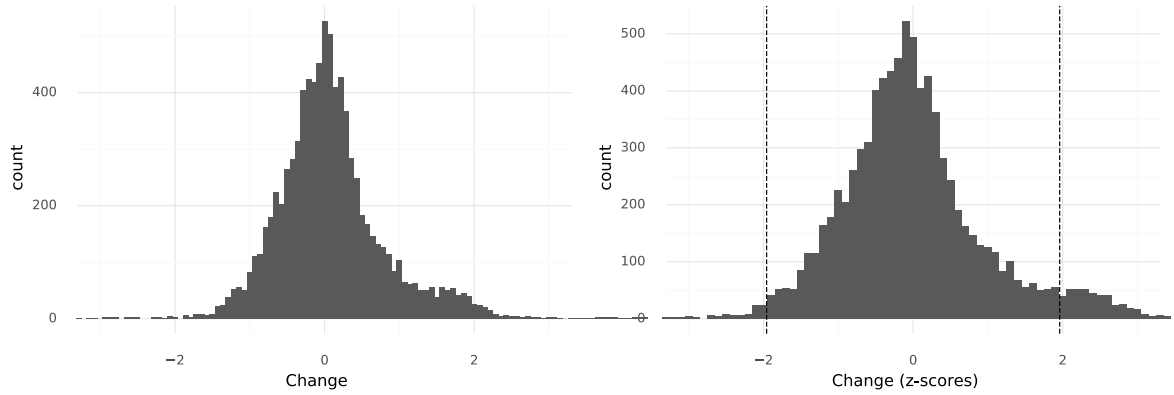

Figure 3.1 Distribution of the differences between the mean intensities in the fasted and fed states. (left) raw values of the difference (mean(fasted) – mean(fed)) (right) normalized values. The normalized values have 0 mean and a standard deviation (SD) of 1. Vertical lines represent 1.96 SD from the mean on either side.

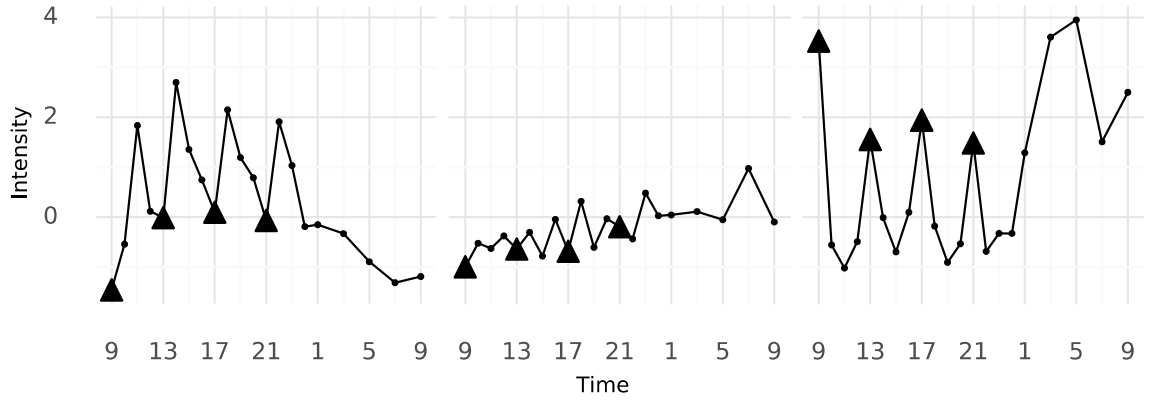

Figure 3.2 24-hour median intensities of representative metabolites showing different types of responses to feeding/fasting. Feeding times are highlighted by triangles.

To formally test whether a metabolite responds to fasting/feeding, the time scale was adjusted to reflect fasting hours. For example, at times 10, 14, 18, and 22 hours, the fasting hour would be 1, since the feeding happened at times 9, 13, 17, and 21 hours. Then the association of the metabolite intensity with fasting hours was measured using a linear mixed effects model, where a random intercept was added for each subject. The p-values for all the metabolites are shown in Figure 3.3. 804 metabolites showed positive significant associations (p-value < 10^{-6}) with fasting hours i.e., the metabolite intensity increased with fasting. To validate, associations for these metabolites were also checked in an independent cohort of 7 subjects (validation cohort) for which we had hourly blood samples. Out of 804, only 510 were found in the validation cohort. For these, the associations

are plotted in Figure 3.4. In total, there were 280 metabolites that showed positive significant (p-value < 10^{-6}) associations with fasting hours in both cohorts.

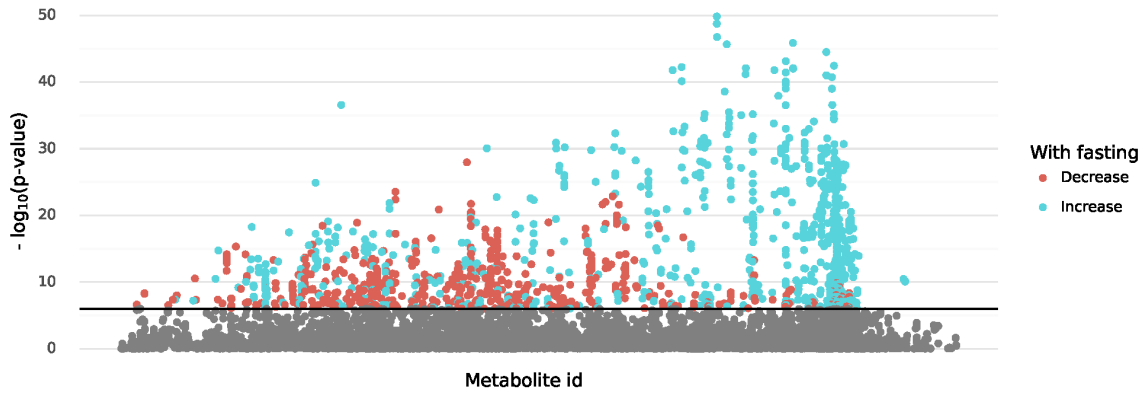

Figure 3.3. p-values for association of metabolite intensities with fasting hours in the linear mixed effect model. Each dot represents a metabolite. Significant associations (p-value < 10^{-6}) are colored based on whether the association with fasting hours is positive or negative.

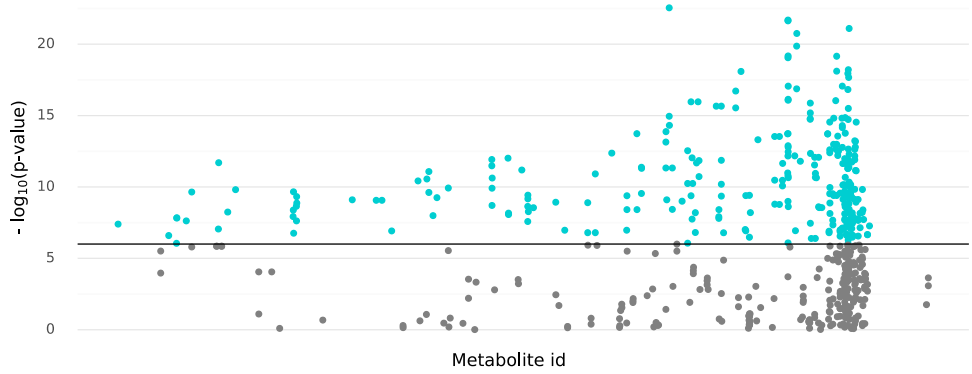

Figure 3.4. p-values for association of metabolite intensities with fasting hours in the validation cohort. Associations are only shown for metabolites that increased with fasting in the discovery cohort.

To better understand the effects of molecules that increase with fasting on human health (fasting molecules), we checked their associations with various obesity and cardio-metabolic phenotypes in an observational cohort of ~8500 subjects with associated clinical information (Cohort 3). These associations are shown in Figure 3.5. Some of these molecules increase with BMI whereas others show higher levels at lower BMIs (negatively correlated with BMI). Interestingly, most of the significant associations of these metabolites with prevalent type 2 diabetes (T2D) and prevalent

coronary heart disease (CHD) were found to be negative (lower levels in subjects with the disease), whereas for HDL and LDL cholesterol, the significant associations were mostly positive.

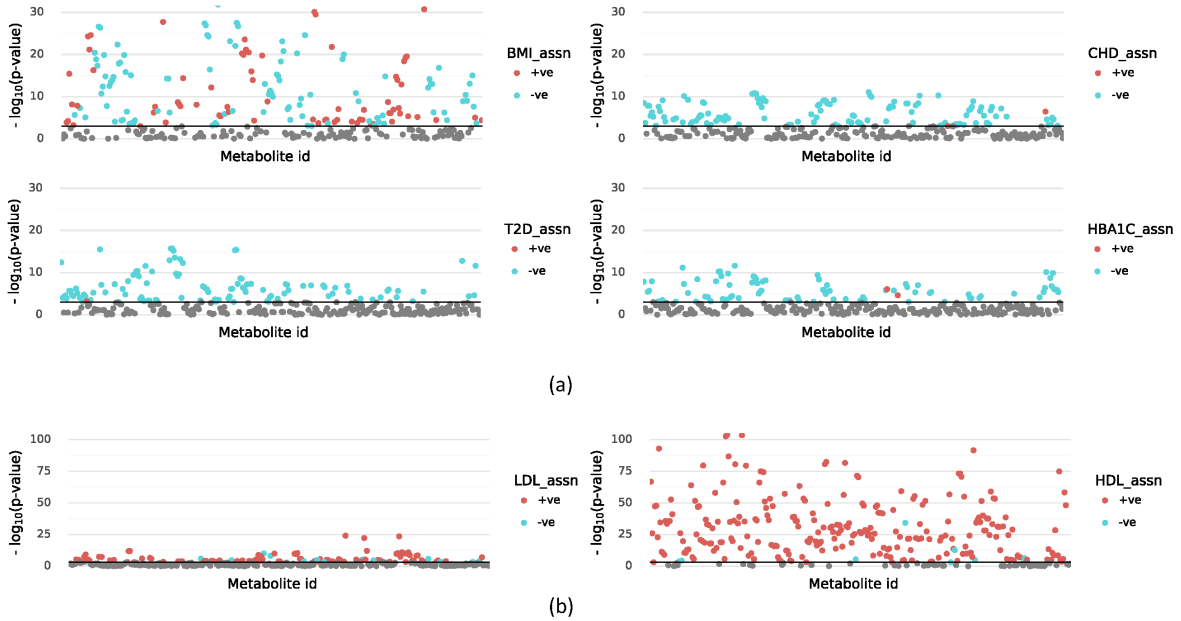

Figure 3.5. Associations of fasting metabolites with various obesity and cardio-metabolic phenotypes. Associations with BMI were checked after adjusting for age and gender and associations for T2D, CHD, HBA1C, LDL, and HDL were checked after adjusting for age, gender, and BMI. Note that the scales on the y-axis are different in subfigures (a) and (b).

3.3 Ongoing and future work

These results beg several follow-up questions. What are the molecular identities of the molecules that increase with fasting? Are any of them FAHFAs? We only had 24 hour sampling data with controlled feeding over a period of 12 hours per day – do these metabolites keep increasing with long term fasting i.e., several days? How do they behave with time-restricted feeding? Are these molecules produced by the adipose tissue? These molecules are associated negatively (protective) with various cardio-metabolic phenotypes including coronary heart disease (CHD) – are these causally protective to CHD i.e., whether the disease is caused as a result of low levels of these molecules? Answers to these questions will open the door to a better understanding of cardio-metabolic health and the effect of fasting on it. However, to answer some of these questions we require more data e.g., data from subjects who followed time-restricted feeding and long-term

fasting. This is ongoing work and results from these analyses will be included in an expanded version of this text, which we intend to publish soon.

3.4 Acknowledgements

Chapter 3 is being prepared for submission for publication of the material. The material is co-authored with members from the Jain Lab, Nallamshetty S., and Rao. R. The dissertation author was one of the primary authors of this material.

Chapter 4

Relationships among PTSD, depression, hostility, and aggression.

4.1 Introduction

Hostility, anger, and aggression are conceptually related but unique constructs. They have been linked to negative outcomes including behavioral and physical health problems, particularly among veterans with posttraumatic stress disorder (PTSD) and/or depression. The health outcomes include acute and chronic pain [65], inflammation [66], headaches [67], lower cognitive function [68], poor sleep quality [69], myocardial infarction and mortality [70, 71], and poor response to mental health treatment [72]. Hostility is defined as an antagonistic attitude or evaluation of others and is associated with feelings of disgust, indignation, and resentment [73]. Anger is an emotional state that consists of feelings that vary in intensity, from a mild irritation or annoyance to fury and rage [74]. Aggression, on the other hand, refers to the behavioral expression of anger that can take the form of physical or verbal acts [75]. Physical aggression can be directed toward self, objects, or others, and verbal aggression can range from shouting angrily to threatening physical violence. Understanding the impact of PTSD and depression on hostility, anger, and verbal and physical

aggression has important clinical implications for assessment and the development of intervention programs for OEF/OIF (Operation Enduring Freedom/Operation Iraqi Freedom) combat veterans.

Recently, there has been an increased interest in the influence of war on aggressive behavior in OEF/OIF service members [76]. PTSD has been associated with physical aggression [77, 78] as well as non-physical forms of aggression [79-81]. Aggression has also been linked to other mental health problems such as depression in veterans and service members [80, 82, 83]. Although both PTSD and depressive symptoms have been found to function almost identically in predicting aggression risk [80, 84], Taft et al. show that when both are considered together, depression ceases to have a significant effect on the presence of aggression. However, depression has been found to partly mediate the relationship between PTSD and two forms of aggression: verbal aggression and aggression toward self [79]. Both PTSD and depression are highly prevalent in service members returning from Afghanistan and Iraq. In a recent study, 15.8% of OEF/OIF veterans screened positive for PTSD [85]. Another study estimated that 13 to 15% of OEF/OIF service members had clinically significant symptoms of depression without PTSD, and 24% had clinically significant levels of comorbid PTSD and depression [86]. Also, OEF/OIF veterans with PTSD have reported higher rates of aggression than veterans without PTSD [87].

A significant limitation in most of the existing aggression research in OEF/OIF and other veterans is the reliance on global measures of aggression [87-90]. It is important to view aggression as a multi-dimensional construct to gain a refined understanding of the types of aggressive behavior, and to improve prediction and measurement of intervention outcomes. The prevalence rates and the risk factors of physical and non-physical aggression differ [91]. Moreover, inclusion of non-physical forms of aggression in a global index can falsely escalate the observed rates of aggression. Therefore, it is necessary to examine aggression as a multi-dimensional construct in order to improve understanding of aggressive behavior in OEF/OIF veterans.

In contrast to aggressive behaviors, fewer studies have examined the relationships between PTSD and hostility in OEF/OIF veterans. Hostility has been associated with both PTSD and

aggression [92-94]. PTSD symptoms are associated with higher hostility scores in both men and women [92]. There is some evidence that hostility partially mediates the association between PTSD and physical aggression [93]. Strong associations have also been found between hostility and depression [68, 95].

Another construct that has been associated with PTSD, depression, hostility and aggression is anger [96, 97]. In order to gain a better understanding of the role of anger and its relationships with other constructs such as PTSD and aggression, anger may be divided into state anger (anger in a given moment) and trait anger (the general propensity to become angry) [98]. Trait anger has been associated with PTSD [99, 100], aggression [78, 99-101], and depression [102]. There is also some evidence that trait anger mediates the relationships between some PTSD clusters and aggression [102].

There has been scant research on hostility, anger and aggression concurrently. Moreover, no studies to date have concurrently examined the direct and indirect effects of PTSD and depression on hostility, anger, and aggression. Given the negative physical, emotional, and psychosocial outcomes associated with these constructs, an examination of the complex relationships between PTSD, depression, hostility, anger, and aggression is important and has the potential to improve the assessment and treatment of OEF/OIF combat veterans.

The aim of the present study was to gain an understanding of the direct and indirect relationships among PTSD, depression, hostility, anger and four types of aggression: (a) verbal aggression, (b) physical aggression toward self, (c) physical aggression toward objects, and (d) physical aggression toward others in a sample of returning OEF/OIF combat veterans. We hypothesized that depression, hostility, and trait anger would mediate the relationship between PTSD and aggression, and that the direct and indirect effects would vary based on type of aggression.

4.2 Methods

4.2.1 Participants

Participants were 175 OEF/OIF combat veterans (95% male, mean age = 30.36 (SD = 8.86)) who were participating in a larger cross-sectional study of genetic factors underlying vulnerability for PTSD. The study excluded participants with a self-reported Axis I disorder diagnosis before deployment (obtained at phone screen), current alcohol dependence, or current drug use. Participants' self-reported ethnicity was 30% Hispanic/Latino, 38% Non-Hispanic/Latino, 32% not reported, and race was 60% white, 6% black, 7% Asian, 2% American Indian, 2% Native Hawaiian/ Pacific Islander, 4% "other", 19% not reported. Recruitment efforts included clinician referrals from VA and Navy clinicians and posting flyers at the VA medical center. The study received local institutional review board approval. All participants provided informed consent before being included in the study.

4.2.2 Procedures

Participants completed self-report questionnaires and a clinical interview to characterize the cohort demographics and to assess study-related constructs. All participants were treated in accordance with the [103].

4.2.3 Measures

Clinician-administered PTSD scale (CAPS). PTSD symptoms were assessed using the CAPS, DSM-IV Version [104], a measure of the frequency and intensity of each of the 17 PTSD symptoms that shows high sensitivity and specificity, high test-retest reliability, and strong convergence with other PTSD self-report measures [105, 106]. In the current sample, internal consistency was excellent (Cronbach's alpha = 0.95).

Beck depression inventory, second edition (BDI-II). The BDI-II [107] contains 21 self-report items that address the cognitive, emotional, and somatic manifestations of depression. Respondents indicate the degree to which they experience symptoms such as hopelessness and

irritability, cognitions such as guilt or feelings of being punished, as well as physical symptoms such as fatigue and weight loss, along a four-point Likert scale ranging from 0 to 3. The items were summed up to construct a total depression score. In the current sample, internal consistency was excellent (Cronbach's alpha = 0.94).

Cook-Medley hostility scale. The Cook-Medley hostility scale [108] is a 50-item hostility scale derived from the Minnesota Multiphasic Personality Inventory (MMPI), that measures different aspects of hostility [109]. For this study, items corresponding to the Cynicism, Hostile Affect, and Hostile Attributions subscales were used. The Aggressive Responding subscale items were not included to avoid any spurious correlations with measures of anger and aggression. In the current sample, internal consistency was good (Cronbach's alpha =0.85).

State-Trait anger expression inventory. Trait anger was measured using the 10-item Trait-Anger (T-Ang) scale of the revised State-Trait anger expression inventory [98], that measures the disposition of someone to express anger with and without provocation. Respondents indicate the frequency with which they experience angry feelings on a 4-point scale, ranging from 1 ("almost never'') to 4 ("almost always"). Scores on individual items were summed up to construct a total T-Ang score. In the current sample, internal consistency was good (Cronbach's alpha = 0.89).

Retrospective overt aggression scale (ROAS). Aggression was measured using the ROAS [110], a retrospective adaptation of the Overt Aggression Scale [111]. Each of the 16 items of the ROAS falls into one of the four subscales - verbal aggression, physical aggression toward objects, physical aggression toward others, and physical aggression toward self, with subscale scoring weighted based on the severity of the aggressive behavior. Respondents indicate the frequency with which they engaged in specific aggressive acts in the past month on a 5-point scale, ranging from 0 ("never / 0 times") to 4 ("always / greater than 10 times"). The ROAS shows excellent inter-rater reliability (r = 0.96), and high intra-class correlations [112, 113]. In the current sample, internal consistency was good (Cronbach's alpha = 0.88).

4.2.4 Data Analysis

Descriptive statistics were computed to examine the sample means and standard deviations on each study measure. Zero-order correlations were computed to examine basic associations between study variables.

Given a dependent variable and its predictor variables, modeling was performed using least-squares regression. The model was constructed as follows: depression was modeled as a function of PTSD, and hostility was modeled as a function of both depression and PTSD. Direct paths were added from PTSD, depression, and hostility to trait anger, and from these four variables to aggression. Four separate analyses were conducted to examine models of the four types of aggression: verbal aggression, physical aggression toward objects, physical aggression toward others, and physical aggression toward self. A direct effect was deemed significant if the corresponding p-value in the linear regression model was smaller than 0.05.

The significance of mediated (or indirect) effects was tested via bootstrapped confidence intervals. Bootstrapping is a non-parametric procedure that, unlike the conventional tests of mediation such as the Sobel test, does not rely on the assumption of normality of the indirect effects' coefficients and generates the distribution empirically by resampling the data with replacement many times [114]. The bootstrapped confidence intervals were generated using 5000 resamples of the data with replacement. An indirect effect was considered significant if the corresponding bootstrapped 95% confidence interval did not contain zero. Age and sex were included as covariates in all direct and indirect effect models.

4.3 Results

4.3.1 Descriptive Statistics and Correlations

The study variable means and standard deviations, and zero-order correlations among them, are presented in Table 4.1. All the variables were positively correlated, $p < 0.001$.

Table 4.1. Scale means, standard deviations (SD), and zero-order correlations. All correlations were significant with $p < 0.001$.

	Mean (SD)	PTSD	TA	Depression	Hostility	VA	POb	POt
PTSD	55.16 (30.44)	-						
TA	20.76 (7.04)	0.59	-					
Depression	15.66 (12.32)	0.74	0.61	-				
Hostility	14.61 (6.13)	0.45	0.57	0.53	-			
VA	8.11 (8.04)	0.52	0.61	0.58	0.43	-		
POb	5.01 (8.41)	0.38	0.49	0.48	0.38	0.65	-	
POt	2.82 (6.98)	0.34	0.38	0.29	0.27	0.48	0.37	-
PS	2.56 (6.09)	0.38	0.40	0.47	0.31	0.48	0.53	0.31

4.3.2 Least Squares Regression Modeling

The standardized direct effects among PTSD, depression, hostility, and trait anger are shown in Figure 4.1. PTSD was a significant predictor of depression ($p < 0.001$). When both PTSD and depression were included as predictors for the criterion variable hostility, the effect of PTSD failed to reach significance ($p = 0.35$), whereas that of depression was significant ($p < 0.001$). Further, each of the variables PTSD, depression, and hostility had a significant effect on trait anger when the other two variables were present in the model.

The direct effects of PTSD, depression, hostility, and trait anger on verbal aggression, physical aggression toward objects, physical aggression toward others, and physical aggression toward self are also reported in Figure 4.1. In the models of verbal aggression and physical aggression toward objects, depression and trait anger had significant direct effects. In contrast, direct effects of PTSD ($p = 0.55$ for verbal aggression, $p = 0.47$ for physical aggression toward objects) and hostility ($p = 0.95$ for verbal aggression, $p = 0.51$ for physical aggression toward objects) failed to reach significance in both models. In the model for physical aggression toward others, trait anger had a significant effect, whereas PTSD ($p = 0.11$), depression ($p = 0.96$), and hostility ($p = 0.73$) did not. However, in the model for physical aggression toward self, only depression had a significant direct effect; the effects of PTSD ($p = 0.93$), hostility ($p = 0.86$) and trait anger ($p = 0.07$) were deemed insignificant. Overall, 56% of the variance in depression, 33% of the variance in hostility, 51% of

the variance in trait anger, 46% of the variance in verbal aggression, 31% of the variance in physical aggression toward objects, 18% of the variance in physical aggression toward others, and 25% of the variance in physical aggression toward self was explained by the models.

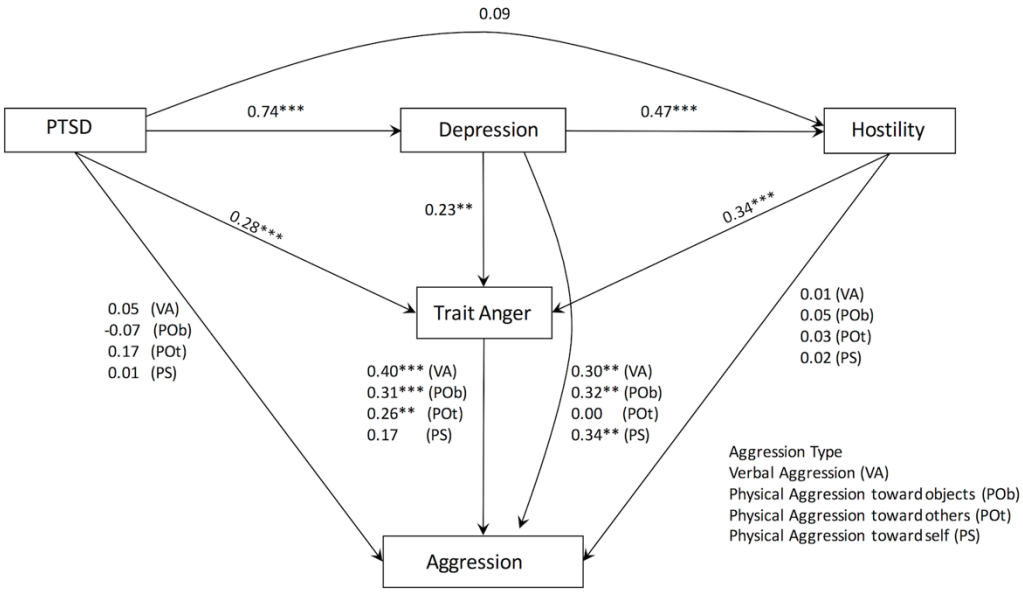

Figure 4.1. Graphical illustration of the direct and indirect effects model of PTSD, depression, hostility, trait anger, and aggression. Numerical values are standardized direct effects. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3.3 Bootstrapped mediation analyses

The results of the bootstrapped mediation analyses are presented in tables 4.2 and 4.3. Depression completely mediated the relationship between PTSD and hostility and moderately mediated the relationship between PTSD and trait anger.

As hypothesized, depression also mediated the associations between PTSD and verbal aggression, physical aggression toward objects, and physical aggression toward self. However, it did not mediate the relationship between PTSD and physical aggression toward others. Trait anger also mediated the associations between PTSD and verbal aggression, physical aggression toward objects, and physical aggression toward others, but not physical aggression toward self.

Furthermore, trait anger completely mediated the association between depression and physical aggression toward others, modestly mediated the relationships between depression and verbal aggression as well as depression and physical aggression toward objects, but did not mediate the relationship between depression and physical aggression toward self.

Trait anger also completely mediated the relationships between hostility and verbal aggression, physical aggression toward objects, and physical aggression toward others. However, for physical aggression toward self, neither hostility nor trait anger had significant direct or indirect effects.

Table 4.2. Standardized direct and indirect effects on depression, hostility, and trait anger (TA). Significant effects (based on 95% confidence intervals) are bolded.

Pathway	Direct Effect (95% CI)	Indirect Effect (95% CI)
PTSD → Depression	**0.74 (0.64 to 0.84)**	
PTSD → Hostility	0.09 (-0.11 to 0.30)	**0.34 (0.21 to 0.48)**
PTSD → Depression → Hostility		**0.34 (0.21 to 0.48)**
Depression → Hostility	**0.47 (0.29 to 0.63)**	
PTSD → TA	**0.28 (0.13 to 0.44)**	**0.32 (0.18 to 0.45)**
PTSD → Depression → TA		**0.17 (0.04 to 0.3)**
PTSD → Hostility → TA		0.03 (-0.04 to 0.1)
PTSD → Depression → Hostility → TA		**0.12 (0.06 to 0.19)**
Depression → TA	**0.23 (0.05 to 0.4)**	**0.16 (0.08 to 0.26)**
Depression → Hostility → TA		**0.16 (0.08 to 0.26)**
Hostility → TA	**0.34 (0.21 to 0.48)**	

Table 4.3. Standardized direct and indirect effects on verbal aggression, physical aggression toward objects, physical aggression toward others, and physical aggression toward self. Significant effects (based on 95% confidence intervals) are bolded.

Pathway	Direct Effect (95% CI)	Indirect Effect (95% CI)
Verbal Aggression (VA)		
PTSD → VA	0.05 (-0.12 to 0.21)	**0.46 (0.28 to 0.67)**
PTSD → Depression → VA		**0.22 (0.06 to 0.4)**
PTSD → Hostility → VA		0.0(-0.02 to 0.02)
PTSD → TA → VA		**0.11 (0.04 to 0.2)**
PTSD → Depression → Hostility → VA		0.0 (-0.05 to 0.04)
PTSD → Depression → TA → VA		**0.07 (0.01 to 0.14)**
PTSD → Hostility → TA → VA		0.01 (-0.02 to 0.04)
PTSD → Depression → Hostility → TA → VA		**0.05 (0.02 to 0.08)**
Depression → VA	**0.30 (0.09 to 0.51)**	**0.16 (0.06 to 0.27)**
Depression → TA → VA		**0.09 (0.02 to 0.19)**
Depression → Hostility → VA		0.00 (-0.06 to 0.06)
Depression → Hostility → TA → VA		**0.06 (0.03 to 0.11)**
Hostility → VA	0.01 (-0.12 to 0.13)	**0.13 (0.07 to 0.22)**
Hostility → TA → VA		**0.13 (0.07 to 0.22)**
TA → VA	**0.40 (0.24 to 0.58)**	
Physical Aggression toward objects (POb)		
PTSD → POb	-0.07 (-0.27 to 0.11)	**0.44 (0.25 to 0.67)**
PTSD → Depression → POb		**0.24 (0.03 to 0.48)**
PTSD → Hostility → POb		0.0 (-0.01 to 0.03)
PTSD → TA → POb		**0.09 (0.03 to 0.17)**
PTSD → Depression → Hostility → POb		0.02 (-0.02 to 0.06)
PTSD → Depression → TA → POb		**0.05 (0.01 to 0.11)**
PTSD → Hostility → TA → POb		0.01 (-0.01 to 0.04)
PTSD → Depression → Hostility → TA → POb		**0.04 (0.01 to 0.07)**
Depression → POb	**0.32 (0.04 to 0.62)**	**0.14 (0.06 to 0.26)**
Depression → TA → POb		**0.07 (0.01 to 0.15)**
Depression → Hostility → POb		0.03 (-0.03 to 0.08)
Depression → Hostility → TA → POb		**0.05 (0.02 to 0.1)**
Hostility → POb	0.05 (-0.06 to 0.17)	**0.10 (0.04 to 0.19)**
Hostility → TA → POb		**0.10 (0.04 to 0.19)**
TA → POb	**0.31 (0.13 to 0.51)**	

Table 4.3. Standardized direct and indirect effects on verbal aggression, physical aggression toward objects, physical aggression toward others, and physical aggression toward self. Significant effects (based on 95% confidence intervals) are bolded.

Physical Aggression toward others (POt)		
PTSD → POt	**0.17 (0.00 to 0.34)**	0.16 (-0.01 to 0.38)
PTSD → Depression → POt		0.0 (-0.14 to 0.15)
PTSD → Hostility → POt		0.0 (-0.02 to 0.03)
PTSD → TA → POt		**0.07 (0.01 to 0.17)**
PTSD → Depression → Hostility → POt		0.01 (-0.04 to 0.07)
PTSD → Depression → TA → POt		**0.04 (0.00 to 0.11)**
PTSD → Hostility → TA → POt		0.01 (-0.01 to 0.03)
PTSD → Depression → Hostility → TA → POt		**0.03 (0.01 to 0.07)**
Depression → POt	0.0 (-0.19 to 0.2)	**0.11 (0.02 to 0.25)**
Depression → TA → POt		**0.06 (0.01 to 0.15)**
Depression → Hostility → POt		0.01 (-0.05 to 0.09)
Depression → Hostility → TA → POt		**0.04 (0.01 to 0.1)**
Hostility → POt	0.03 (-0.12 to 0.18)	**0.09 (0.02 to 0.18)**
Hostility → TA → POt		**0.09 (0.02 to 0.18)**
TA → POt	**0.26 (0.06 to 0.51)**	
Physical Aggression toward self (PS)		
PTSD →PS	0.01 (-0.19 to 0.17)	**0.37 (0.10 to 0.70)**
PTSD →Depression → PS		**0.25 (0.04 to 0.53)**
PTSD → Hostility → PS		0.0 (-0.02 to 0.02)
PTSD → TA → PS		0.05 (-0.01 to 0.13)
PTSD → Depression → Hostility → PS		0.01 (-0.05 to 0.05)
PTSD → Depression → TA → PS		0.03 (-0.01 to 0.09)
PTSD → Hostility →TA → PS		0.01 (-0.01 to 0.03)
PTSD → Depression → Hostility → TA → PS		0.02 (-0.01 to 0.06)
Depression → PS	**0.34 (0.06 to 0.67)**	**0.07 (0.00 to 0.18)**
Depression → TA → PS		0.04 (-0.01 to 0.12)
Depression → Hostility → PS		0.01 (-0.06 to 0.07)
Depression → Hostility → TA → PS		0.03 (-0.01 to 0.08)
Hostility → PS	0.02 (-0.13 to 0.15)	0.06 (-0.02 to 0.15)
Hostility → TA → PS		0.06 (-0.02 to 0.15)
TA → PS	0.18 (-0.05 to 0.42)	

4.4 Discussion

This was the first study to investigate the direct and indirect relationships among PTSD, depression, hostility, anger and four types of aggression: (a) verbal aggression, (b) physical aggression toward self, (c) physical aggression toward objects, and (d) physical aggression toward others in a sample of returning OEF/OIF combat veterans. PTSD, depression, hostility, and anger

behaved differently while predicting different forms of aggression, corroborating the recognition of aggression as a multi-dimensional construct. Based on linear regression analysis, only depression had a significant direct effect on physical aggression toward self, whereas in the models for verbal aggression and physical aggression toward objects, depression and trait anger had significant effects. In contrast, in the model of physical aggression toward others, only trait anger had a significant direct effect.

Although the bootstrapped analysis was only used to test the significance of indirect effects, the 95% bootstrapped confidence intervals for direct effects are also reported in tables 4.2 and 4.3 for the sake of completeness. It is important to mention that the bootstrapped confidence interval for the direct effect of PTSD on physical aggression toward others (*95% CI*: 0.003-0.335) does not contain zero even though the effect is non-significant based on linear regression analysis ($p = 0.11$). For all other direct effects, there is consistency between the significance as determined by least squares regression and the 95% bootstrapped confidence interval not containing zero.

In support of our study hypothesis, depression mediated the associations between PTSD and physical aggression toward self, verbal aggression, and physical aggression toward objects, when hostility and trait anger were in the model. This is partly consistent with the findings of one study [79], where depressive symptoms partly mediated the relationship between PTSD and two forms of aggression: verbal aggression and physical aggression toward self. In the study, however, the authors only considered PTSD, depression, and aggression, whereas in this study, hostility and trait anger were also in the model. Contrary to our hypothesis, depression did not mediate the association between PTSD and physical aggression toward others. This is also consistent with the findings of two studies [79, 84], that observed that although depression and PTSD both uniquely predicted physical aggression, when considered together, depression ceased to have a significant effect.

To the best of our knowledge, no studies have investigated whether trait anger mediates the associations between PTSD and different forms of aggression, although one study [102] showed that trait anger mediated the relationships between some PTSD clusters and general aggression, which

was measured only using three items related to verbal and physical aggression. In this study, as hypothesized, trait anger mediated the associations between PTSD and verbal aggression, physical aggression toward objects, and physical aggression toward others, when depression and hostility were in the model. It did not, however, mediate the relationship between PTSD and physical aggression toward self. This study, therefore, provides a unique contribution to the literature by providing the relationships between PTSD, trait anger, and different forms of aggression.

In contrast to trait anger, the indirect effect of PTSD on aggression via hostility alone was not significant for any type of aggression. This is contrary to a study [93], where hostility partially mediated the association between PTSD and physical aggression, measured by items corresponding mostly to physical aggression toward others. One reason why the findings differ could be the inclusion of depression and trait anger in our models of aggression. In our study, PTSD had no significant direct effect on hostility when depression was in the model, and hostility had no significant direct effect on physical aggression towards others when trait anger was in the model.

Current results highlight that PTSD is not the overall direct contributor to different forms of aggression, and clearly show the role of depression and trait anger. Depression symptoms explain part of the relationships between PTSD and verbal aggression, physical aggression toward objects, and physical aggression toward self, and trait anger explains part of the relationships between PTSD and verbal aggression, physical aggression toward objects, and physical aggression toward others. Concurrent PTSD symptoms and higher trait anger in veterans may warrant close monitoring of veterans for being physically aggressive toward others. On the other hand, veterans with high depressive symptoms should be closely monitored for self-harm. Although evidence-based treatments for PTSD can help reduce depressive symptoms [115], explicitly treating depression among OEF/OIF veterans with PTSD may reduce acts of physical aggression toward self. Similarly, explicitly addressing trait anger along with PTSD symptoms among OEF/OIF veterans may help reduce incidents of physical aggression toward others.

The findings of this study should be viewed in the context of some limitations. First, our sample is predominantly male and white, which might limit generalizability of the findings. However, white males constitute majority of the US veterans [116], suggesting that the findings are applicable to the general population of interest. Second, participants may not have accurately reported aggressive acts on the retrospective self-report aggression scale. Finally, these data were cross-sectional, and hence, any purported causal pathways must be cautiously considered. Although future research on the relationships between PTSD, depression, hostility, anger, and aggression is needed for validation, our findings suggest that clinicians working with veterans should consider a multifaceted approach to treatment that not only addresses PTSD, but also depression and trait anger.

4.5 Acknowledgements

Chapter 4 is adapted from Bhardwaj, V., Angkaw, A. C., Franceschetti, M., Rao, R., & Baker, D. G. (2019). Direct and indirect relationships among posttraumatic stress disorder, depression, hostility, anger, and verbal and physical aggression in returning veterans. Aggressive behavior, 45(4), 417-426. The dissertation author was the primary author of this paper.

Appendix A

Supplement Notes on Chapter 2

Exact algorithm for solving the String Reconstruction Problem

Greedy Algorithm

MINING-D Parameters

Defining Relative Positions

Removing Unidirectional Extensions

Immunosequencing Datasets

Benchmarking MINING-D against IgScout

Novel Variations

Finding D genes in Whole Genome Sequencing Data

D Gene Usage

Overused D Genes

Highly Used D Genes in Non-human Datasets

Benchmarking MINING-D on simulated CDR3s

Benchmarking MINING-D on TCR datasets

Non-genomic insertions in naive and cord blood Rep-Seq datasets

A.1 Exact algorithm for solving the String Reconstruction Problem

It is easy to see that $P(C|s)$ is maximized by one of the modified strings. This observation leads to a brute-force algorithm for solving the String Reconstruction Problem (with complexity $O(|s|*N^2)$) that simply computes $P(C|s)$ for each of the N modified strings. Below, we describe a $O(|s|*N)$ algorithm for solving this problem that is linear in the input size.

We assume for simplicity that all modified strings are different. This is not a strict assumption as one can always add special symbols to distinguish all strings. We denote $f(j) = \log(|\mathcal{A}|^{j+1} - 1)$ and search for a string that maximizes $\sum_{i=1}^{N} f(m_i)$. We denote a t-symbol prefix (t-*prefix*) of a string c as c^t and the set of all t-prefixes of strings from C as C^t. Given a string s and an integer t, we say that a string c is t-*similar* to s if t-prefixes of s and c coincide. The number of strings in C that are t-similar to s is denoted as $sim_t(C,s)$. Given a string s,

$$\text{score}(C^t|s^t) = \text{score}(C^{t-1}|s^{t-1}) + sim_t(C,s) \times \log\left(\frac{|\mathcal{A}|^{t+1} - 1}{|\mathcal{A}|^t - 1}\right). \quad (7)$$

We use this recurrence to efficiently compute score($C|s$) for each string s from C using dynamic programming. We construct a *trie* of all strings in C [117]. Each vertex in the trie is a t-prefix s^t of a string from C, and we recursively compute score ($C^t|s^t$) in each vertex of the trie using the above recurrence (assuming that the score of the root is $N \times \log(|\mathcal{A}| - 1)$). The optimal string is the string corresponding to the leaf node with the maximum score (Figure A.1). All scores can be computed by a single Depth First Search, assuming that all values $sim_t(C,s)$ are computed during the construction of the trie.

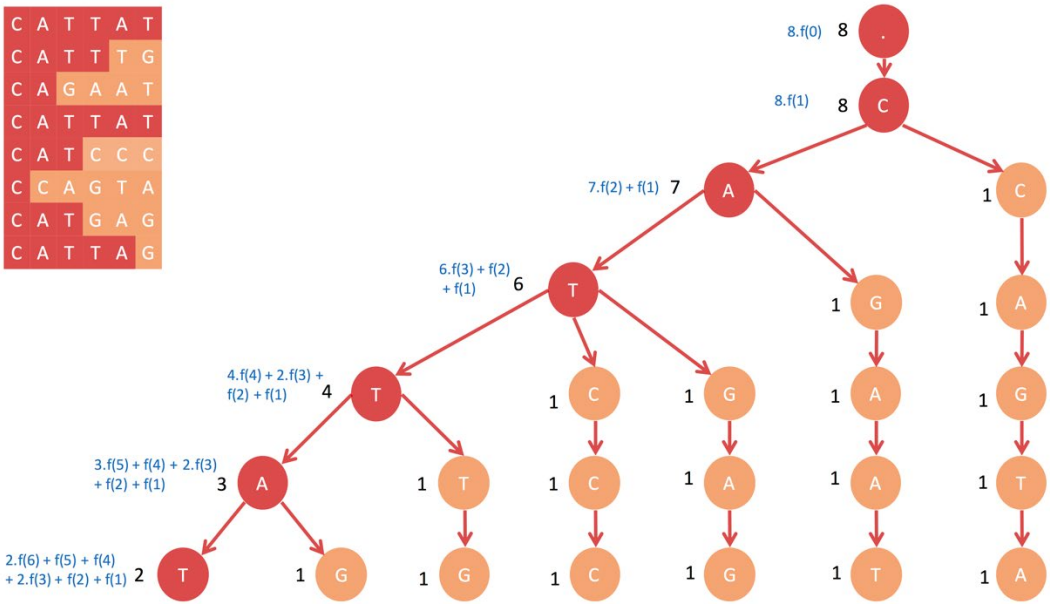

Figure A.1 Illustration of the algorithm for solving the String Reconstruction Problem. The set of modified strings is shown on the left, and their trie is shown on the right. The string associated with each vertex is the one that is formed by traversing from the root node to the vertex. The number of leaves under each vertex is shown on the left. The scores for all vertices in the path from the root node to the leaf node with the maximum score are shown in blue. The leaf CATTAT is the optimal seed string.

A.2 Greedy Algorithm

The pseudocode of the greedy algorithm is as follows:

greedy_string *(C):*
    $S_g \leftarrow emptystring()$
    $for\ j\ in\ 1\ to\ q$ *:*
        $E(j) \leftarrow$ *most abundant symbol at position* j *in the strings in C*
        $S_g \leftarrow S_g + E(j)$
        $C \leftarrow$ *set of all strings* c_i *in C such that* $c_i[1{:}\,j\,] = S_g$
    $return\ S_g$

Figure A.2. Pseudocode of the greedy algorithm.

A.3 MINING-D Parameters

The most important parameter of the MINING-D algorithm is *m,* the number of seed *k-mers.* The default value of *m* should be different across species, since different numbers of D genes take part in the recombination process in each species. To decide on the default *m* for each species, we applied MINING-D to all datasets with different values of *m*. The results are shown in Table A.1.

Based on the results in the table, we chose the following as the default values: human (m = 600), mouse (m= 300), rat (m = 300), rhesus macaque (m = 600), Bactrian camel (m = 300), and rabbit (m = 100).

The p-value threshold was chosen to be 10^{-36}. This value achieves 80% power from the test with a sample size of 2000 when the effect size (deviation from uniform distribution) is medium, according to the definition of the medium effect for chi-squared test. Having a strict (very low) threshold on the p-value may lead to some missing nucleotides on the sides of the genes, but since we are also doing genomic validation, the whole gene can be recovered from the genomic reads. On the other hand, high p-value threshold will not only lead to extra nucleotides on the sides, it will also cause more extensions to be made from a single k-mer, leading to more false positives. As another test, we also tried to extend the known human IMGT genes in Healthy Human CDR3 datasets using this threshold. 95% of the time, no extension was made to any gene.

Table A.1. Information about inferred D genes. *m* denotes the number of seed 10-mers. The number of novel genes and variations validated using genomic data are also shown.

Species - Individuals	IMGT Database	*m*	# Inferred genes including variations	# Inferred genes in IMGT	# Novel variations (validated)	# Novel genes (validated)
Healthy Humans 20	Human	1000	42	25	12 (2)	5 (0)
		600	38	25	8 (2)	5 (0)
		300	27	20	5 (2)	2 (0)
		100	15	12	3 (1)	-
Untreated + Immunized Mouse 27	Mouse	1000	35	18	12 (1)	5 (0)
		600	29	18	9 (1)	2 (0)
		300	24	18	5 (1)	1 (0)
		100	17	15	2 (1)	-
Immunized Wistar Rat 1	Rats	1000	27	13	11 (3)	4 (0)
		600	20	13	6 (3)	1 (0)
		300	16	12	4 (3)	-
		100	13	9	4 (3)	-
Rhesus macaque - 7	Crab-eating macaque	1000	25	17	6 (6)	2 (2)
		600	25	17	6 (6)	2 (2)
		300	24	16	6 (6)	2 (2)
		100	14	10	2 (2)	2 (2)
Bactrian Camels 3	Alpaca	1000	24	2	18 (8)	4 (0)
		600	19	2	15 (8)	2 (0)
		300	13	1	12 (8)	-
		100	10	1	9 (7)	-
Immunized New Zealand Rabbit 1	Rabbit	1000	73	3	57 (3)	13 (0)
		600	53	3	39 (3)	11 (0)
		300	34	3	25 (3)	6 (0)
		100	18	3	13 (3)	2 (0)

A.4 Defining Relative Positions

Looking at the relative positions of the extensions of *k*-mers has some advantages over looking at the relative positions of the *k*-mers. Since a relatively short *k*-mer can be a part of two of

the three types of V, D, and J genes, the mean relative position among all the CDR3s of which such a *k*-mer is a substring can be misleading. Moreover, even if the *k*-mer is a substring of only one gene, the relative position of the extension gives a better estimate of the position of the CDR3 part of which the *k*-mer is a substring as illustrated in Figure A.3 and Figure A.4.

Figure A.3. The relative position of a 10-mer in a CDR3. The red, green, and blue colors represent parts of the V, D, and J segments in a CDR3 sequence. The relative position of the 10-mer CGAAATACTA is 0.32, whereas the relative position of its potential extension in red is 0.04.

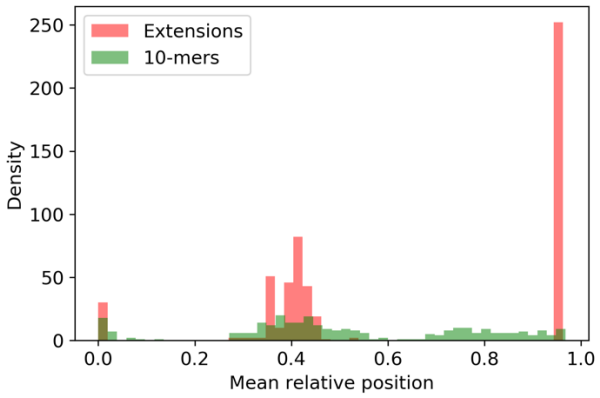

Figure A.4. The mean relative positions of the abundant seed 10-mers (in green) and their extensions (in red) in the MOUSE dataset. The relative positions of the extensions form three clusters, each corresponding to one of the V, D, and J genes.

A.5 Removing Unidirectional Extensions

Not all the unique extensions in the central cluster correspond to different D genes. Some of them are multiple reconstructions of the same D gene and are very similar to each other in the sense that they differ from each other by only a few nucleotides only at the edges. Most of them can be eliminated by making the observation that a highly abundant *k*-mer that the algorithm starts with might not always be, as a whole, a substring of a D gene. For example, the *k*-mer shown in Figure A.5 can be among the highly abundant *k*-mers chosen to extend if the D gene shown in (a) is represented highly in the CDR3 sequences. When extended, it only extends to the right as shown in

(c), retaining the random insertions in the *k*-mer. We can eliminate such *unidirectional* extensions because we expect some of the central *k*-mers of the D gene to also be among the highly abundant 10-mers. Such *k*-mers will be extended in both directions (*bidirectional* extensions), and by eliminating the unidirectional extensions, we reduce the number of reconstructions per D gene.

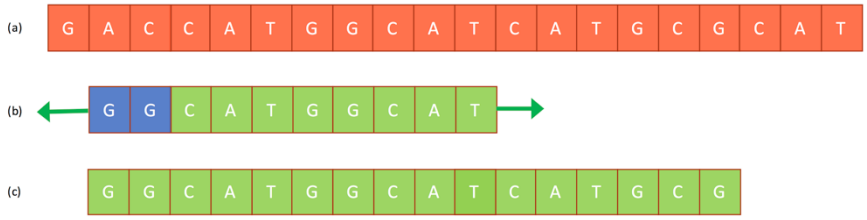

Figure A.5. A highly abundant 10-mer (b) that is formed by random insertions (two nucleotides in the beginning) and 8 nucleotides from a highly abundant D gene (a). Since this 10-mer was not a substring of the D gene, its extension (c) is also not a substring of the D gene.

Formally, let the number of nucleotides added to the left and right of the *k*-mer be N_L and N_R, respectively. We put the following constraint on N_L and N_R:

$$\frac{|N_L - N_R|}{\max(N_L, N_R)} \leq \alpha,$$

where α is a parameter of the algorithm. We used $\alpha = 0.5$. The possible values of N_L and N_R with $\alpha = 0.5$ are shown in Table A.2.

Table A.2. Possible values of N_L and N_R with the constraint when $\alpha = 0.5$.

N_L	N_R
1	1,2
2	1,2,3,4
3	2,3,4,5,6
4	2,3,4,5,6,7,8
5	3,4,5,6,7,8,9,10

A.6 Immunosequencing Datasets

Summaries of all the human and non-human immunosequencing datasets analyzed in this study are shown in Table A.3 and Table A.4, respectively.

Table A.3. Summary of human immunosequencing datasets analyzed in the study. ASC refers to antibody secreting cells.

Name	# Individuals	Tissue	Cell Types	Isotypes	NCBI Project	# Datasets
Allergy Patients	6	PBMC, Bone Marrow	Unsorted	NA	PRJEB18926	24
Flu Vaccination	8	PBMC	Unsorted, Memory, resting memory, HA+/- memory, naive, ASC	NA	PRJNA324093	95
	3	PBMC	Unsorted	NA	PRJNA349143	18
Healthy	3	PBMC	Unsorted	IgG, IgM	PRJNA430091	28
Cord Blood	5	PBMC, Cord Blood	Unsorted	NA	PRJNA393446	6
Intestinal Repertoire	7	Ileum Mucosa, Colon Mucosa	Memory, Plasma	IgA, IgM	PRJNA355402	35
Multiple Sclerosis	4	Brain lesion, Cervical lymph node, Choroid plexus, Pia mater	Unsorted	NA	PRJNA248475	32
Hepatitis B (a)	9	PBMC	Unsorted, HBsAg+ and HLA-DR+ plasma cells	IgG	PRJNA308566	142
Hepatitis B (b)	9	PBMC	Unsorted, HBsAg+ and HLA-DR+ plasma cells	IgG, IgM	PRJNA308641	107

Table A.4. Summary of non-human immunosequencing datasets analyzed in the study.

Species	Strains	Health Status	# Individuals	Tissue	Cell Types	Isotypes	Project	# Datasets
Mouse	C57BL/6J, Balb/c, Pet shop	Untreated, Antigen-immunized	27	Spleen, Bone marrow	pre-B cell, long lived plasma cell, naive B cell,	NA	PRJEB18631	71
Macaque	Rhesus macaques of Indian and Chinese origin	Healthy	7	PBMC	unsorted	IgM	PRJEB15295	7
Camel	Bactrian	Healthy	3	PBMC	PBMC	VH, VHH	PRJNA321369	6
Rat	Wistar	Immunized	10	Spleen	unsorted	NA	PRJNA386462	10
Rabbit	New Zealand white rabbit	Sequentially immunized	3	PBMC, Spleen	unsorted	NA	PRJNA355270	7

A.7 Benchmarking MINING-D against IgScout

We compared the results of IgScout and MINING-D on all datasets from the project PRJEB18926. The results are shown in Figure A.6. A gene is said to be present in a dataset if at least one variation of the gene is found in the dataset and missing otherwise. In most datasets, both IgScout and MINING-D miss three D genes with very low usage (IGHD1-14, IGHD1-20, IGHD6-25) and a very short IGHD7-27 gene (11 nt). These D genes are also reported as missing in multiple studies on analyzing the usage of D genes [43, 118-120]. While IgScout also misses three more short D genes with low usage i.e., IGHD1-1, IGHD4-4, IGHD1-7, MINING-D infers these genes for some individuals.

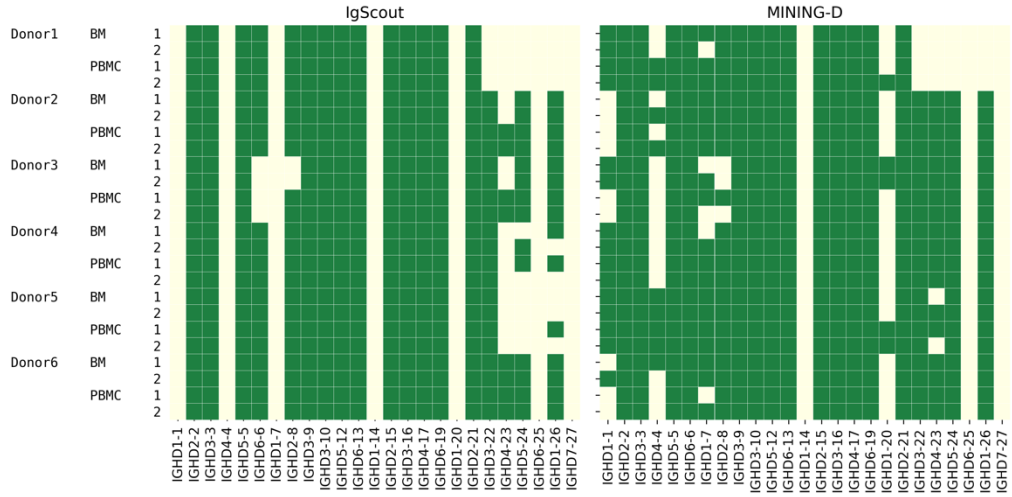

Figure A.6. Results of IgScout (left) and MINING-D (right) on datasets from the project PRJEB18926. All genes that were found in a dataset are shown in dark green, whereas the missing genes in datasets are denoted by light green. Missing inferences for genes IGHD3-22 through IGHD7-27 in Donor 1 indicate a potential deletion polymorphism in the IGHD locus.

We also compared the MINING-D and IgScout results on non-human datasets. Tables A.5 through Table A.7 compare the results of IgScout and MINING-D on ten Mouse datasets (4 Balb/c mice, 4 C57BL/6 mice, and 2 pets), all Rat datasets, and all Camel datasets. Figure A.7 presents the distributions of missing and extra nucleotide bases in the inferred genes (as compared to the IMGT genes for all mouse datasets) for both MINING-D and IgScout.

Table A.5. **Comparison of IMGT genes inferred by IgScout and MINING-D in Mouse datasets.** The gene IGHD2-10*01 (the only gene inferred by IgScout but missed by MINING-D) and the gene IGHD2-1*01 (inferred by MINING-D but missed by IgScout) only differ at the first position.

Strain	Dataset	Inferred IMGT D genes		
		Both	IgScout only	MINING-D only
Balb/c	ERR1759659	IGHD1-1*0 IGHD1-2*01 IGHD2-1*01 IGHD2-14*01 IGHD2-3*01 IGHD2-4*01 IGHD3-2*01 IGHD4-1*01 IGHD2-10*01	-	IGHD2-10*02 IGHD2-2*01
Balb/c	ERR1759660	IGHD1-1*01 IGHD1-2*01 IGHD2-14*01 IGHD2-3*01 IGHD2-4*01 IGHD3-2*01 IGHD4-1*01 IGHD2-10*01	-	IGHD2-10*02 IGHD2-2*01
Balb/c	ERR1759661	IGHD1-1*01 IGHD1-2*01 IGHD2-14*01 IGHD2-3*01 IGHD2-4*01 IGHD3-2*01 IGHD4-1*01 IGHD2-10*01	-	IGHD2-10*02 IGHD2-2*01
Balb/c	ERR1759662	IGHD1-1*01 IGHD1-2*01 IGHD2-1*01 IGHD2-14*01 IGHD2-3*01 IGHD2-4*01 IGHD3-2*01 IGHD4-1*01	-	IGHD2-10*02 IGHD2-2*01
C57BL/6	ERR1759665	IGHD1-1*01 IGHD2-1*01 IGHD2-3*01 IGHD2-4*01 IGHD2-5*01 IGHD3-2*02 IGHD4-1*01	-	IGHD2-2*01
C57BL/6	ERR1759668	IGHD1-1*01 IGHD2-1*01 IGHD2-3*01 IGHD2-4*01 IGHD2-5*01 IGHD4-1*01 IGHD3-2*02	-	IGHD2-2*01
C57BL/6	ERR1759671	IGHD1-1*01 IGHD2-1*01 IGHD2-3*01 IGHD2-4*01 IGHD2-5*01 IGHD3-2*02 IGHD4-1*01	-	IGHD2-2*01
C57BL/6	ERR1759674	IGHD1-1*01 IGHD2-1*01 IGHD2-3*01 IGHD2-4*01 IGHD2-5*01 IGHD4-1*01 IGHD3-2*02	-	IGHD2-2*01
Pet	ERR1759679	IGHD1-1*01 IGHD1-2*01 IGHD2-14*01 IGHD2-3*01 IGHD2-4*01 IGHD3-2*02 IGHD4-1*01	IGHD2-10*01	IGHD2-1*01 IGHD2-2*01 IGHD2-5*01 IGHD3-2*01
Pet	ERR1759680	IGHD1-1*01 IGHD1-2*01 IGHD2-14*01 IGHD2-3*01 IGHD2-4*01 IGHD3-2*01 IGHD4-1*01 IGHD2-10*01	-	IGHD2-2*01

Table A.6. Comparison of IMGT genes inferred by IgScout and MINING-D from Rat datasets.

Dataset	Inferred IMGT D genes		
	Both	IgScout only	MINING-D only
SRR5534359	IGHD1-10*01 IGHD1-11*01 IGHD1-12*03 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD4-3*01 IGHD5-1*01	-	IGHD1-1*01 IGHD1-12*02
SRR5534360	IGHD1-1*01 IGHD1-10*01 IGHD1-11*01 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD1-9*01 IGHD4-3*01	IGHD5-1*01	IGHD1-12*02 IGHD1-12*03 IGHD1-8*01
SRR5534361	IGHD1-11*01 IGHD1-2*01 IGHD1-6*01 IGHD4-3*01	-	IGHD1-1*01 IGHD1-10*01 IGHD1-12*02 IGHD1-12*03 IGHD1-4*01 IGHD1-5*01
SRR5534362	IGHD1-1*01 IGHD1-10*01 IGHD1-11*01 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD4-3*01	IGHD5-1*01	IGHD1-12*02 IGHD1-12*03
SRR5534363	IGHD1-10*01 IGHD1-11*01 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD4-3*01 IGHD5-1*01	-	IGHD1-1*01 IGHD1-12*02 IGHD1-12*03
SRR5534364	IGHD1-1*01 IGHD1-10*01 IGHD1-11*01 IGHD1-12*02 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD1-9*01 IGHD4-3*01 IGHD5-1*01	-	IGHD1-12*03 IGHD1-8*01
SRR5534365	IGHD1-1*01 IGHD1-10*01 IGHD1-11*01 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD1-9*01 IGHD4-3*01 IGHD5-1*01	-	IGHD1-12*02 IGHD1-12*03 IGHD1-7*01 IGHD1-8*01
SRR5534366	IGHD1-10*01 IGHD1-11*01 IGHD1-12*02 IGHD1-12*03 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD1-8*01 IGHD1-9*01 IGHD4-3*01 IGHD5-1*01	-	IGHD1-1*01
SRR5534367	IGHD1-10*01 IGHD1-11*01 IGHD1-12*02 IGHD1-2*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD1-8*01 IGHD1-9*01 IGHD4-3*01	IGHD5-1*01	IGHD1-1*01 IGHD1-12*03
SRR5534368	IGHD1-1*01 IGHD1-10*01 IGHD1-11*01 IGHD1-4*01 IGHD1-5*01 IGHD1-6*01 IGHD1-9*01 IGHD4-3*01 IGHD5-1*01	-	IGHD1-12*02 IGHD1-12*03 IGHD1-2*01 IGHD1-8*01

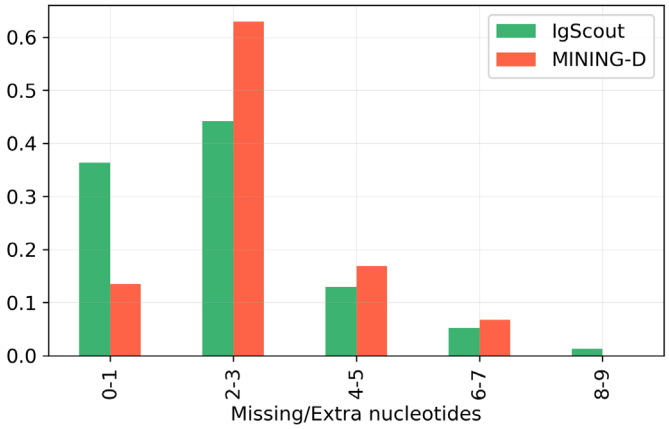

Figure A.7. Distribution of missing or extra nucleotide bases in the inferred genes as compared to the IMGT genesfor all mouse datasets shown in Table E. Only genes that were inferred by both IgScout and MINING-D were included in the comparison.

Table A.7. Comparison of genes inferred by IgScout and MINING-D from the Camel datasets. M-I denotes that a gene was inferred by both MINING-D and IgScout, whereas M denotes that the gene was inferred by MINING-D only. IMGT genes in this table refer to the IMGT alpaca genes. Only genes that were validated using genomic reads are included in this comparison.

Gene (Alpaca)	Variant	1 VH	1 VHH	2 VH	2 VHH	3 VH	3 VHH
IGHD6*01	N_Var (IGHD6*01) - 2	M-I	M-I	M-I	M-I	M-I	M-I
IGHD2*01	N_Var (IGHD2*01) - 1	M-I	M-I	-	M-I	M-I	M-I
	N_Var (IGHD2*01) - 0	-	-	M-I	-	-	-
IGHD3*01	N_Var (IGHD3*01) - 1	M-I	M-I	**M**	M-I	M-I	M-I
	N_Var (IGHD3*01) - 0	-	-	M-I	-	-	-
IGHD4*01	N_Var (IGHD4*01) - 1	M-I	M-I	**M**	-	**M**	-
	N_Var (IGHD4*01) - 0	-	-	M-I	-	-	-
IGHD5*01	N_Var (IGHD5*01) - 0	**M**	**M**	-	**M**	**M**	**M**
	IGHD5*01	-	-	**M**	-	-	-

A.8 Novel Variations

All the variations found using MINING-D for humans, camels, rhesus macaques, mice, rats, and rabbits are shown in Table A.8. The polymorphisms in the genes validated using genomic data are highlighted.

Table A.8. All Inferred novel variations. Variations that were validated using genomic data are shown by highlighting the polymorphisms.

Human

IGHD3-10*01

```
Original    GTATTACTATGGTTCGGGGAGTTATTATAAC
N_Var-3     GTATTACTATGGTTCAGGGAGTTATTATAAC
N_Var-2     --------ATGGTTCGGGGAcTTAT------
N_Var-1     ---------TGGTTCGGGGAaTTA-------
N_Var-0     ---------TGaTTCGGGGAGTT--------
```

IGHD3-22*01

```
Original    GTATTACTATGATAGTAGTGGTTATTACTAC
N_Var-1     --ATTACTATGATAcTAGTGG----------
N_Var-0     -------TATGATAGcAGTGGT---------
```

IGHD2-2*01

```
Original    AGGATATTGTAGTAGTACCAGCTGCTATGCC
N_Var-0     AGGATATTGTAtTAGTACCAGCTGCTAT---
```

IGHD3-16*02

```
Original    GTATTATGATTACGTTTGGGGGAGTTATCGTTATACC
N_Var-0     ---TTATGATTACATTTGGGGGAGTTATCGTTAT---
```

Camel

IGHD3*01 (Alpaca)

```
Original   GTATTACTACTGCTCAGGCTATGGGTGTTATGAC
N_Var-1    ----GACTGCTATTCAGGCTCTTGGTGTTATG--
N_Var-0    ---TGACTACTGTTCAGGCTCTTGGTGT------
```

IGHD2*01 (Alpaca)

```
Original    ACATACTATAGTGGTAGTTACTACTACACC
N_Var-1     --ATATTGTAGTGGTGGTTACTGCTAC---
N_Var-0     -CATACTATAGTGGTGGTTACTAC------
```

IGHD4*01 (Alpaca)

```
Original    TTACTATAGCGACTATGAC
N_Var-1     CTACTATAGCGACTATG--
N_Var-0     -TACTATAACGAATATG--
```

IGHD6*01 (Alpaca)

```
Original    GTACGGTAGTAGCTGGTAC----
N_Var-4     --ACGGTgGTAGtTGGT------
N_Var-3     ---CGGTgGTAGgTGGTggctgg
N_Var-2     GTACGGTGGTAGCTGGTAC----
N_Var-1     ---CGGTgGTAcCTGGT------
N_Var-0     --ACGGTgGTAaCTGG-------
```

IGHD5*01 (Alpaca)

```
Original    AGACTACGGGTTGGGGTAC
N_Var-0     ----TATGGGTT-GGGTAC
```

Rhesus Macaque

IGHD1S39*01

```
Original    GGTATAGTGGGAACTACAAC
N_Var-0     -----AGTGGGAGCTAC---
```

IGHD3S18*01

```
Original    GTACTGGGGTGATTATTATGAC
N_Var-0     --ACTGGAGTGATTATTA----
```

IGHD5S3*01

```
Original    GTGGATACAGTGGGTACAGTTAC
N_Var-0     ---GATACAGCGGGTACAGT---
```

IGHD2S11*01

```
Original    AGAATATTGTAGTAGTACTTACTGCTCCTCC
N_Var-0     -----ATTGTAGTGGTACTTACTGCT-----
```

IGHD2S17*01

```
Original    AGAATACTGTACTGGTAGTGGTTGCTATGCC
N_Var-0     ----TACTGTACTGGTAGTGGTTGCTAC---
```

IGHD3S23*01

```
Original    GTATTACTATGATAGTGGTTATTACACCCACAGCGT
N_Var-0     ---TTACTATGGTAGTGGTTATTAC----------
```

Mouse

IGHD1-1*01

```
Original    TTTATTACTACGGTAGTAGCTAC-
N_Var-3     ------ACgACGGTAGTAGC----
N_Var-2     -TTATTACTACGGTAGTAGagggg
N_Var-1     ---ATTACTgCGGTAGTAGCTAC-
N_Var-0     TTTATTACTACGATGGTAGCTACg
```

IGHD2-4*01

```
Original    TCTACTATGATTACGAC---
N_Var-0     ------ggGATTACGACagg
```

Rat

IGHD1-3*01

```
Original    TTTTTAACTATGGTAGCTAC
N_Var-0     -TTTTAACTACGGTAGCTAC
```

IGHD1-9*01

```
Original    TACATACTATGGGTATAACTAC-
N_Var-1     --CATACTACGGGTATACCTACg
N_Var-0     --CATACTAcGGGTATAACTAC-
```

IGHD1-12*02

```
Original    TTTATTACTATGATGGTAGTTATTACTAC-
N_Var-0     -TTATTACTATGATGGTACTTATTACTACg
```

Rabbit

IGHD6-1*01

```
Original    --------------GTTACTATAGTTATGGTTATGCTTATGCTACC
N_Var-7     ---------------TTAtgATgGTTATGGTTATGgTa--------
N_Var-6     tta------tagtgGTTAtggTgGTTATGcTTATG-----------
N_Var-5     ---------------TggtTATgGTgATGGTTATGCT----------
N_Var-4     TTACTATACTTATGGTTATGCTGGTTATGCTTATGCTAC-------
N_Var-3     TTA------TGCTGGTTATGCTGGTTATGGTTATGCTAC-------
N_Var-2     ------------tgGTTAtggTgGTTATGGTTATG-----------
N_Var-1     --------------------ATAcTTATGGTTATGgTggT------
N_Var-0     --------------------ATAGTTATGGTTATGgTg--------
```

IGHD1-1*01

```
Original    GCATATACTAGTAGTAGTGGTTATTATATAC
N_Var-2     GCATATGCTAGTAGTAGTGGTTATTAT----
N_Var-1     -----------TgGTAGTGGTTATTAT----
N_Var-0     -------------GTAGTGGTggTTAT----
```

IGHD2-1*01

```
Original    TAGCTACGATGACTATGGTGATTAC-
N_Var-0     ---------TGAtTATGGTGgTTAtg
```

IGHD8-1*01

```
Original    GTTATGCTGGTAGTAGTTATTATACC
N_Var-0     -TTATGCTGGTgaTgGTTATg-----
```

A.9 Finding D genes in Whole Genome Sequencing data

Tables A.9 through Table A.13 show the number of WGS reads confirming both novel and known variations of D genes and demonstrate that novel and known D genes have similar numbers of supporting reads in the selected WGS datasets.

Table A.9. Number of genomic reads containing exact occurrences of known and novel allelic variants of human genes IGHD3-10, IGHD3-16, and IGHD3-22 in five datasets containing reads with the novel allelic variants. Information on the number of datasets that a gene is present in and the number of reads containing the gene for some other genes - IGHD2-2*01: 36(4-54), IGHD3-3*01: 30(6-45), IGHD3-22*01: 40(5-29).

Dataset	IGHD3-10			IGHD3-16			IGHD3-22
	*01	*02	N_Var (IGHD3-10*01)-3	*01	*02	N_Var (IGHD3-16*02)-0	
SRR6435661	8	-	14	-	42	30	16
SRR6435676	19	-	12	-	77	58	28
SRR6435686	14	-	8	-	20	32	12
SRR6435691	4	-	12	-	35	32	13
SRR6435692	6	-	12	-	35	36	15

Table A.10. Number of genomic reads containing exact occurrences of known and novel allelic variants of macaque genes. IGHD3S18*01 has the same sequence as IGHD3S29*01 that results in higher coverage than the novel allelic variant. Similarly, the coverage for IGHD5S3*01 is higher than the novel variant because it has the same sequence as IGHD5S25*01.

Gene	Type	SRR7865780	SRR7865781	SRR7865793	SRR7865795
IGHD1S39	IMGT	26	16	11	0
	Novel	0	0	0	18
IGHD3S18	IMGT	69	45	54	29
	Novel	0	21	0	8
IGHD5-S3	IMGT	74	58	73	41
	Novel	28	22	22	13
IGHD2S11	IMGT	24	27	36	7
	Novel	0	0	0	6
IGHD2S17	IMGT	0	0	0	9
	Novel	30	12	32	8
IGHD3S23	IMGT	0	9	8	11
	Novel	24	12	17	0
N_Gene-0		27	9	21	9
N_Gene-1		0	8	0	9

Table A.11. Number of genomic reads containing exact occurrences of known and novel allelic variants of rat genes.

Gene	IGHD1-3		IGHD1-9		IGHD1-12			
	IMGT	Novel	IMGT	Novel	IMGT*01	IMGT*02	IMGT*03	Novel
SRR7503107	5	11	-	2	-	-	4	8
SRR7503108	8	6	-	4	-	-	2	1
SRR7503109	8	7	-	5	-	-	4	8
SRR7503110	12	7	-	4	-	-	6	8
SRR7503111	5	6	-	5	-	-	4	9
SRR7503112	6	2	-	2	-	-	7	2
SRR7503113	7	5	-	3	-	-	4	7
SRR7503114	12	4	-	6	-	-	7	6
SRR7503115	9	18	-	8	-	-	17	9
SRR7503116	1	5	-	5	-	-	2	3

Table A.12. Number of reads containing exact occurrences for known and novel allelic variants of rabbit genes.

Gene	Type	Datasets present in (#Reads)
IGHD6-1	IMGT	8 (1-5)
	N_Var (IGHD6-1*01)-3	19 (1-9)
	N_Var (IGHD6-1*01)-4	11 (1-6)
IGHD1-1	IMGT	6 (2-3)
	N_Var (IGHD1-1*01)-2	23 (1-14)

Table A.13. Number of genomic reads containing exact occurrences of inferred camel genes. The IMGT genes here correspond to alpaca genes.

Gene	Type	SRR19472 39	SRR19472 40	SRR19472 41	SRR19472 42	SRR19472 43	SRR19472 44	SRR19472 45
IGHD3*01	IMGT	-	-	-	-	-	-	-
	N_Var - 0	-	-	-	-	2	6	-
	N_Var - 1	16	6	7	12	1	8	11
IGHD2*01	IMGT	-	-	-	-	-	-	-
	N_Var - 0	-	-	-	-	2	2	-
	N_Var - 1	10	11	5	17	4	-	5
IGHD4*01	IMGT	-	-	-	-	-	-	-
	N_Var - 0	-	-	-	-	5	4	-
	N_Var - 1	14	11	15	10	-	7	10
IGHD5*01	IMGT	-	-	-	-	-	-	-
	N_Var - 0	13	9	2	6	5	0	3
IGHD6*01	IMGT	-	-	-	-	-	-	-
	N_Var - 2	11	9	21	11	5	14	14

A.10 D Gene Usage

Usage of D genes in the Flu Vaccination dataset. To analyze the usage of the D genes in different types of cells (including hemagglutinin-positive (HA+) and HA- activated B cells, antibody secreting cells, memory cells, and naive cells) from PBMCs at different times after flu vaccination, we used 95 datasets from the NCBI project PRJNA324093. 55.3% of CDR3s were traceable on average in all the datasets. The D gene usage profiles are very different in HA+ cells and other cells for almost all individuals suggesting the usage of specific D genes for HA+ clones in those individuals (Figure A.8). Interestingly, the overused D genes are not the same across individuals. For instance, for individual 7, genes IGHD2-21*02 and IGHD4-17*01 are overused, and for individual 6, only the gene IGHD4-17*01 is overused.

Usage of D genes in the Multiple Sclerosis dataset. 45.2% of CDR3s on average were traceable in each dataset. The usage of D genes across datasets from tissues such as brain lesion, cervical lymph node, choroid plexus, and pia mater is shown in Figure A.9. The results suggest that the usage of genes is different in different tissues from the same individual. For instance, for individual M5, IGHD1-26*01 and IGHD3-3*01 are overused in choroid plexus, whereas only IGHD3-3*01 is overused in brain lesion compared to other tissues.

Usage of D genes in the Intestinal Repertoire dataset. We analyzed the usage of D genes in datasets corresponding to memory and plasma cells, IgA and IgM isotypes from ileum and colon tissues from 4 individuals, and naive cells from ileum from 3 individuals (Figure A.10). 43.5% of CDR3s on average were traceable in each dataset. For IgM naive cells (ileum mucosa), the number of traceable CDR3s was 71.42% on average, whereas for memory and plasma cells from the same tissue, it was 43.25% and 43.12%, respectively. The D gene IGHD3-3*01 was used significantly less in plasma and memory cells from both tissues compared to naive cells from ileum and PBMCs from healthy individuals (Figure 4). Similarly, the gene IGHD6-6*01 seems to be under-used in plasma and memory cells from the ileum tissue compared to naive cells. Subtle differences can also be found

among the usage between different isotypes from the same individual's tissue, e.g., genes IGHD2-21*02, IGHD2-8*01, IGHD3-16*02, IGHD5-5*01/IGHD5-18*01, and IGHD7-27*01 are presented more in the IgM isotype than the IgA isotype in the colon tissue from individual 0.

Usage of D genes in the Hepatitis B Vaccination dataset. To study the usage of D genes in HbsAg+ B cells and HLA-DR+ plasma cells, we analyzed datasets corresponding to individuals who received a Hepatitis B vaccination. 51.3% of CDR3s on average were traceable in each dataset. IgM and IgG datasets had 65.4% and 45.9% traceable CDR3s on average, respectively. The usage of genes is shown in Figure A.11. Differences in the usage profiles can be seen among HbsAg+ B cells, HLA-DR+ plasma cells, and PBMCs from the same individual for most of the individuals. For instance, for individual 7, IGHD2-15*01 is under-used in both HbsAg+ B cells and HLA-DR+ plasma cells compared to PBMCs, whereas genes IGHD4-17*01 and IGHD4-23*01 are overused. The gene IGHD3-22*01 is unpresented and the genes IGHD5-5*01/IGHD5-18*01 and IGHD7-27*01 are overused in HLA-DR+ plasma cells compared to HbsAg+ B cells. For individual 2, as another example, the genes IGHD3-22*01, IGHD3-3*01, and IGHD6-13*01 do not appear to be presented in the CDR3s from HLA-DR+ plasma cells, although they are presented in both the PBMCs and HbsAg+ B cells from the same individual. Similarly, differences between profiles can be found for all individuals, although there does not appear to be a strong pattern across individuals, suggesting that the response is highly personalized and might depend upon other factors.

Usage of D genes in Cord Blood dataset. 48.9% of CDR3s were traceable on average in the PBMC datasets, whereas 71.6% of the datasets were traceable in the Cord Blood datasets (Figure A.12). Supplemental Note: "Non-genomic insertions in naive and cord blood Rep-Seq datasets" shows that the Cord Blood datasets are characterized by smaller number of VD and DJ insertions compared to the naïve datasets.

Usage of D genes in mice datasets. Figure A.13 shows usage of various known and novel genes/variations in different datasets corresponding to different strain, cell type, and tissue from mice.

Usage of D genes in the Rhesus macaque datasets. 52.6% of CDR3s on average were traceable in each dataset. The usage of the IMGT genes and the validated novel genes and variants of known genes is shown in Figure A.14.

Usage of D genes in the Camel datasets. 31.7% of CDR3s on average were traceable in each dataset. Although the small sample size (n = 3) limits generalizability, the low number of traceable CDR3s could be due to high level of hypermutation within the CDR3 region as compared to other species. Since there is no IMGT database for camels, we used the alpaca IMGT database as a reference to analyze the usage. The usage of these genes and the validated novel variants of these genes is shown in Figure A.15. It can be seen that the D gene usage profiles are very different for the VH and the VHH isotypes within individuals, especially for individuals 2 and 3.

Usage of D genes in the Rat datasets. 54.3% of CDR3s on average were traceable in each dataset. The usage of the IMGT genes and the validated novel variants is shown in Figure A.16. Genes belonging to the IGHD2 and IGHD3 families were underutilized as compared to other gene families, and the novel variants were among the genes that were utilized in most of the datasets. There is no clear distinction between the usage profiles between HuD and DNP immunized rats. This could be due to one or more of the following reasons: (a) the CDR3s here are from unsorted cells from spleen and not antigen specific cells; (b) the usage profiles of individuals might not be identical before immunization, hence masking the pattern if there was any.

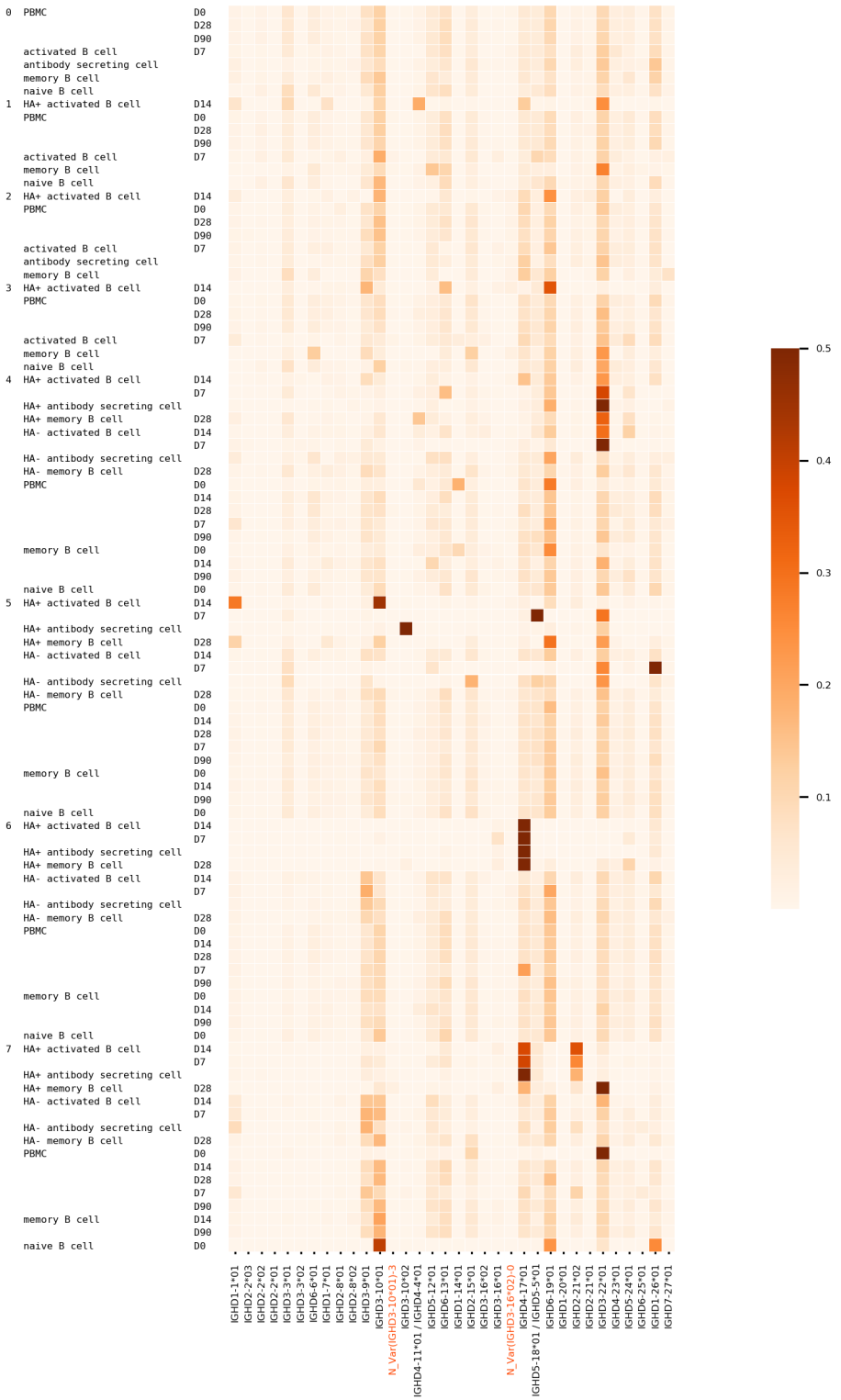

Figure A.8. Usage of IMGT and novel variations of IGHD genes in various datasets corresponding to flu vaccination. The columns on the left represent the individual, the cell type, and the time point (day after vaccination).

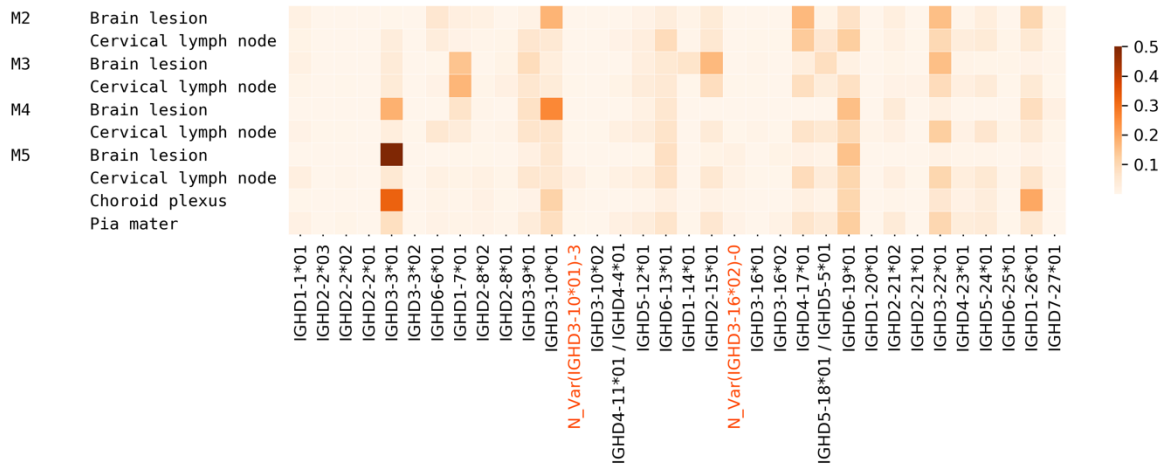

Figure A.9. Usage of various known and novel genes in various datasets corresponding to different tissues in Multiple Sclerosis patients. The columns on the left represent the individual and the tissue, respectively.

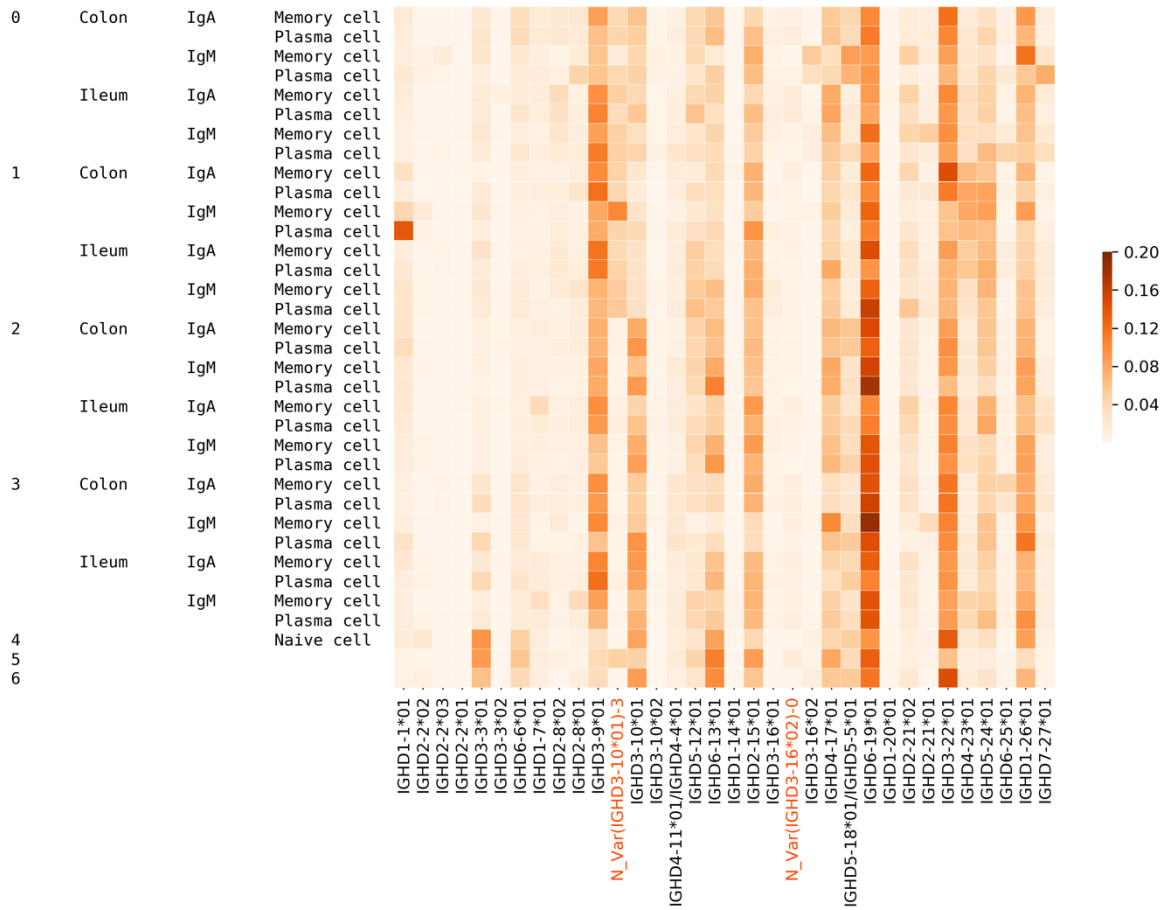

Figure A.10. Usage of various known and novel genes in various datasets corresponding to human intestinal antibodies. The columns on the left represent the individual, tissue, isotype, and cell type, respectively.

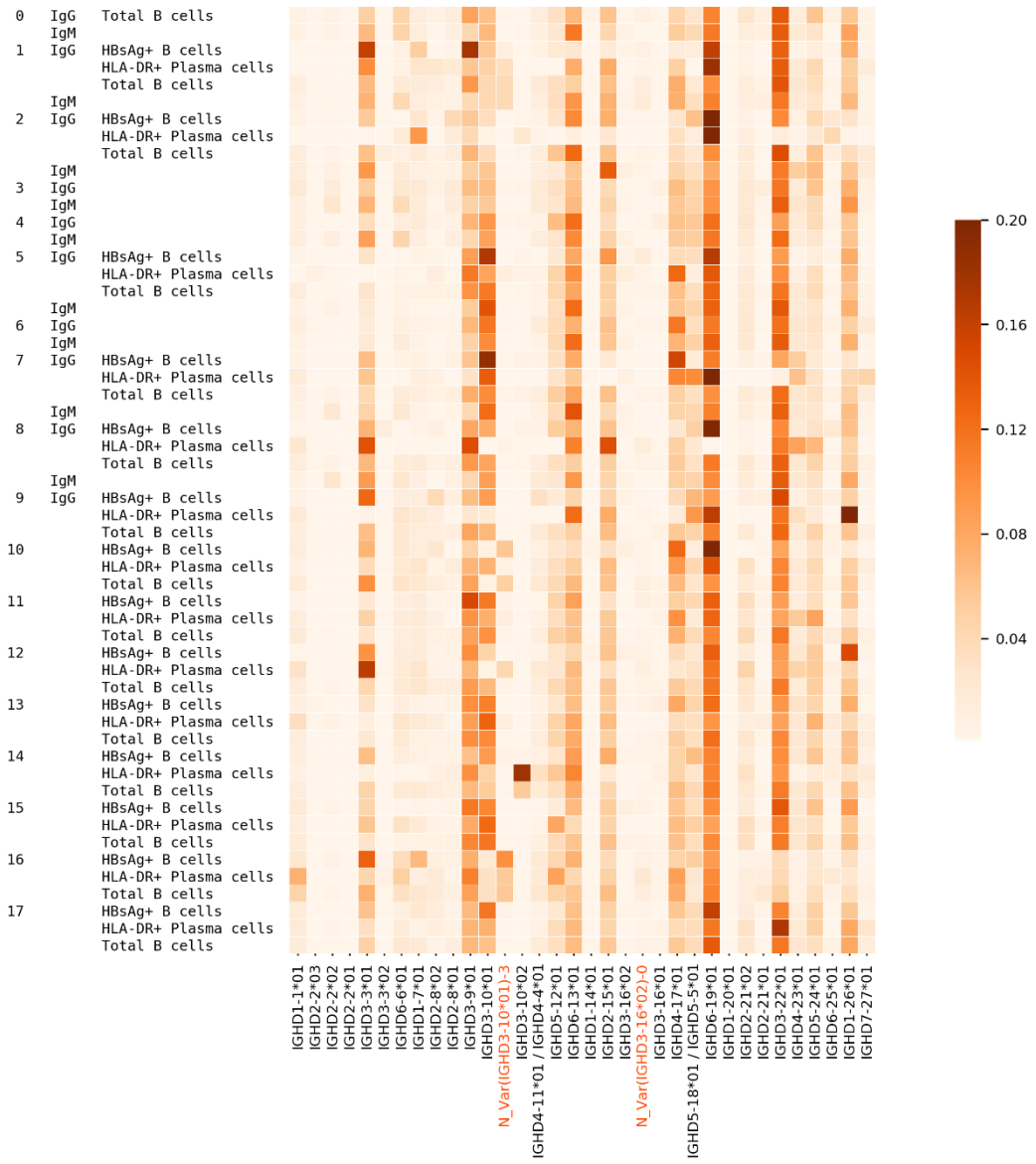

Figure A.11. Usage of various known and novel genes in different datasets corresponding to different cell types and isotypes corresponding to human subjects with hepatitis B vaccination.

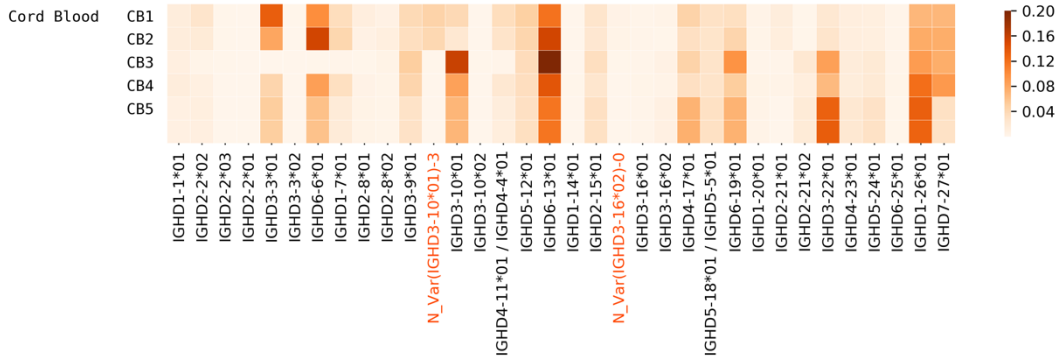

Figure A.12. Usage of various known and novel genes in cord blood datasets.

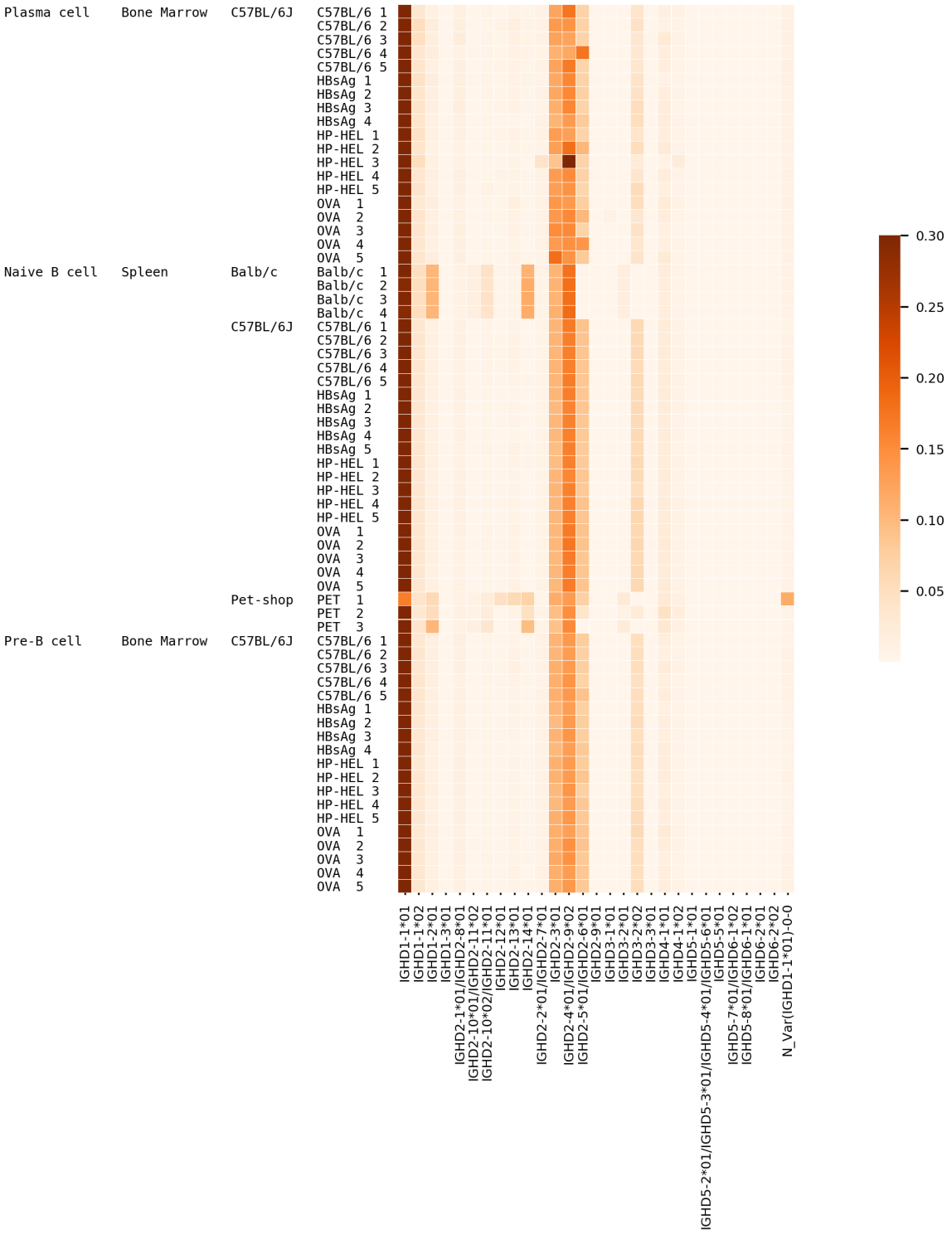

Figure A.13. Usage of various known and novel genes/variations in different datasets corresponding to different strain, cell type, and tissue from mice. Columns on the left represent cell type, tissue, strain, and individual, respectively. OVA, HP-HEL, and HBsAg in the right most column represent the C57BL/6J mice immunized with OVA, HP-HEL, and HBsAg, respectively. For example, OVA 3 represents the C57BL/6J mouse number 3 that was immunized with OVA.

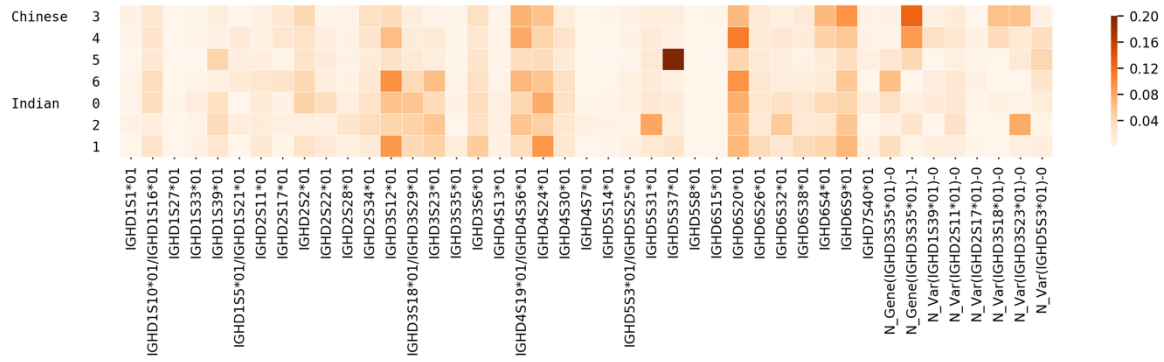

Figure A.14. Usage of known and novel genes in the Rhesus Macaque datasets. The novel genes and variations are shown on the right.

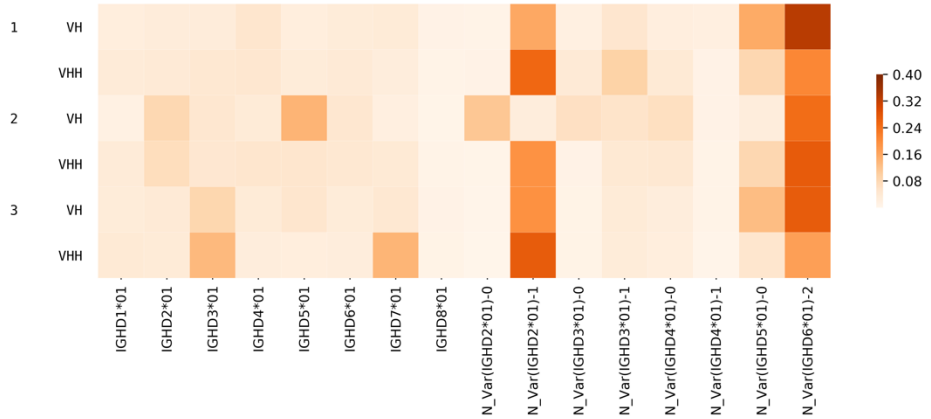

Figure A.15. Usage of known and novel genes in the Camel datasets.

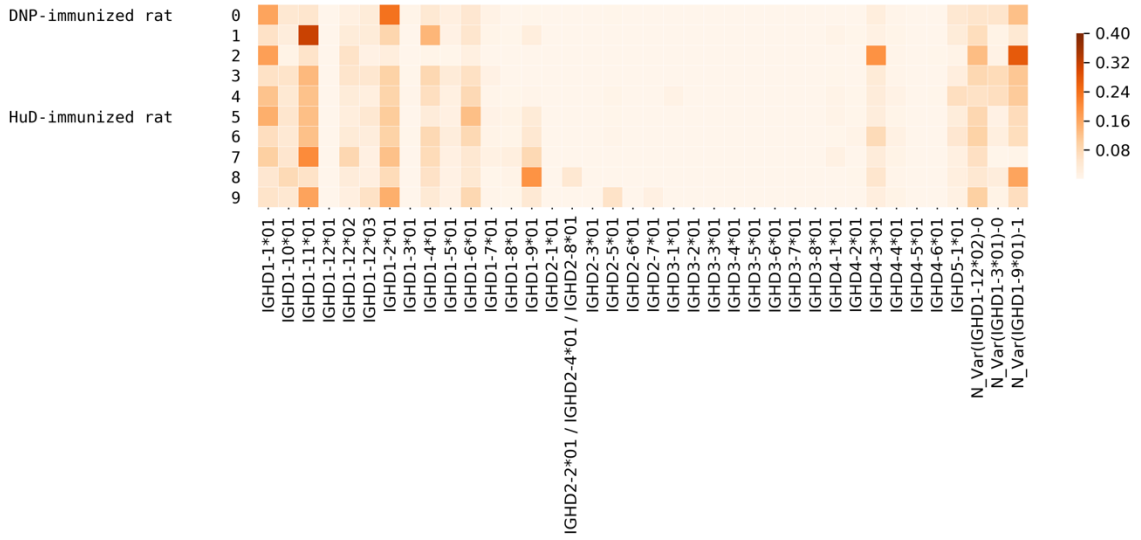

Figure A.16. Usage of D genes in the Rat datasets.

A.11 Overused D genes

Tables A.14 through Table A.17 show overused genes in different datasets.

Table A.14. Overused genes in the Multiple Sclerosis datasets.

Gene	Donor	Tissue	Over-usage
IGHD1-7*01	M3	Brain Lesion	2.7x
		Cervical Lymph Node	3.1x
IGHD3-3*01	M4	Brain Lesion	2.6x
	M5	Brain Lesion	8.2x
		Choroid Plexus	4.7x

Table A.15. Overused genes in the Intestinal Repertoire datasets.

Gene	Donor	Tissue	Isotype	Cell type	Overusage
IGHD1-1*01	1	Colon	IgM	Memory B cell	2.0x
				Plasma cell	6.6x
IGHD2-21*01	0	Ileum	IgM	Memory B cell	5.0x
	3	Colon			3.6x
IGHD2-8*02	0	Ileum	IgA	Memory B cell	2.7x
				Plasma cell	2.2x
			IgM	Memory B cell	2.0x
IGHD3-16*02	0	Colon	IgM	Memory B cell	5.4x
				Plasma cell	3.3x
IGHD4-23*01	1	Colon	IgA	Plasma cell	2.3x
			IgM	Memory B cell	2.3x
IGHD6-25*01	0	Colon	IgM	Plasma cell	2.4x
		Ileum		Plasma cell	5.4x
				Memory B cell	2.0x
	3	Colon	IgA	Memory B cell	5.2x
				Plasma cell	3.1x

Table A.16. Overused genes in the Hepatitis B vaccination datasets

Gene	Isotype	Individual	Cell type	Over usage
IGHD1-1*01	IgG	16	HLA-DR+ Plasma cells	3.5x
			Total B cells	2.3x
IGHD3-10*02		14	HLA-DR+ Plasma cells	43.8x
			Total B cells	13.3x
		5	HLA-DR+ Plasma cells	2.1x
		2		5.7x
		8	HBsAg+ B cells	2.1x
IGHD3-3*01		12	HLA-DR+ Plasma cells	2.4x
		1	HBsAg+ B cells	2.9x
		8	HLA-DR+ Plasma cells	2.0x
IGHD6-25*01		13	HLA-DR+ Plasma cells	2.2x
		16		2.4x
		2		4.8x
		8	HBsAg+ B cells	2.4x
		2		2.6x

Table A.17. Overused genes in the Cord Blood datasets.

Gene	Individual	Overusage
IGHD7-27*01	CB1	3.4x
	CB2	3.8x
	CB3	3.8x
	CB4	4.6x

A.12 Highly Used D Genes in Non-human Datasets

To find the genes with the highest usage among the datasets of a species, we picked the top 3 genes from each dataset. A gene is said to be *highly used* in all datasets from a species if it is one of the top 3 genes in at least 3 datasets. We found 3 highly used D genes for camels, 3 for macaques, and 4 for rats (Table A.18 and Figure A.17).

Table A.18. Highly used D genes in the Camel, Macaque, and Rat datasets. Genes shown here are among the top 3 genes in terms of usage proportion in the number of datasets shown in the right column.

Species (Total datasets)	D Gene	Datasets
Camel (6)	N_Var (IGHD2*01)-1 N_Var (IGHD5*01)-0 N_Var (IGHD6*01)-2	5 3 6
Macaques (7)	IGHD3S12*01 IGHD4S19*01/IGHD4S36*01 IGHD6S20*01	3 3 5
Rats (10)	IGHD1-1*01 IGHD1-11*01 IGHD1-2*01 N_Var (IGHD1-9*01)-1	5 7 6 5

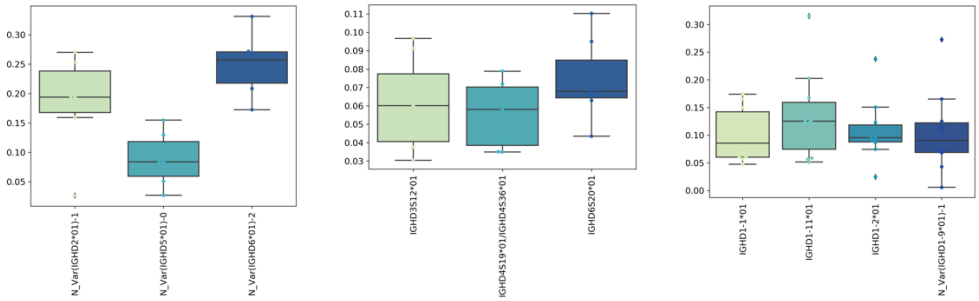

Figure A.17. Usage proportion of highly used genes in the Camel (left), Macaque (middle), and Rat (right) datasets.

A.13 Benchmarking MINING-D on simulated CDR3s

We simulated 250,000 mutation-free CDR3s using the human D genes listed in the IMGT database (except for IGHD1-14*01, IGHD4-23*01, IGHD5-24*01) and IgSimulator tool [121]. We then generated four mutated versions of each of these CDR3s using mutation rates equal to 0.01, 0.05, 0.1, and 0.2. In total, we had one unmutated and four mutated datasets resulting in the average number of SHMs per CDR3 equal to 0, 0.7, 3.7, 7.4, and 14.8, respectively. For datasets with mutability < 0.1, MINING-D inferred all genes except for IGHD7-27*01 and one of the allelic variants of the gene IGHD2-2. There were no missing or additional nucleotide bases in the inferred D genes as compared to the D genes used for simulating the CDR3s. The missed gene IGHD7-27*01 is the shortest human D gene (11 nucleotides) that cannot be inferred using the default value of k (k=10) for MINING-D. MINING-D inferred only one of the allelic variants IGHD2-2*01 and IGHD2-2*03 since they differ only at the first base as shown below.

```
IGHD2-2*01        AGGATATTGTAGTAGTACCAGCTGCTATGCC
IGHD2-2*02        AGGATATTGTAGTAGTACCAGCTGCTATACC
IGHD2-2*03        TGGATATTGTAGTAGTACCAGCTGCTATGCC
```

In the dataset with the mutation rate 0.1, in addition to missing the gene IGHD7-27*01, MINING-D inferred only one sequence for the three allelic variants of IGHD2-2 shown above. As the mutation rate increased to 0.2, a similar pattern was observed for IGHD3-16. Moreover, only the first sequence was inferred for genes IGHD1-20*01 and IGHD1-7*01 shown below.

```
IGHD1-20*01       GGTATAACTGGAACGAC
IGHD1-7*01        GGTATAACTGGAACTAC
```

For the dataset with the mutation rate 0.2, multiple partial sequences were inferred per gene for some of the genes. However, there were no falsely inferred genes. For example, the following two sequences were inferred for the gene IGHD2-15*01

```
IGHD2-15*01       AGGATATTGTAGTGGTGGTAGCTGCTACTCC
```

```
Inferred-1          GATATTGTAGTGGTGGTAG

Inferred-2                GTAGTGGTGGTAGCTGCTAC
```

Table A.19. Results of MINING-D on simulated datasets. X*** denotes that only a single allelic variant sequence was inferred among multiple allelic variants of a D gene X. X/Y*** denotes that only a single sequence was inferred for two D genes X and Y.

Mutation rate	Avg #SHMs per CDR3	#D genes used in the simulation	#D genes inferred by MINING-D	Missing D genes	Mean #missing/ extra nucleotides per inferred gene	Falsely inferred genes
0	0	29	27	IGHD2-2*03 IGHD7-27*01	0	0
0.01	0.74	29	27	IGHD2-2*01 IGHD7-27*01	0	0
0.05	3.70	29	28	IGHD2-2*01 IGHD7-27*01	0	1
0.1	7.40	29	26	IGHD7-27*01 IGHD2-2***	1.04	0
0.2	14.79	29	33	IGHD7-27*01 IGHD2-2*** IGHD3-16*** IGHD1-20/1-7***	8.21	0

A.14 Benchmarking MINING-D on TCR datasets

We downloaded ten TRB cell datasets corresponding to 7 individuals from the immuneACCESS database (Table A.20). Each dataset consists of short sequences (~100 nt) fully covering CDR3s and partially covering V and J genes. Since our preprocessing step is not designed for such short sequences, we skipped CDR3 search step and used original sequences as an input for MINING-D. Information about the D genes in the TRBD locus and the datasets they were inferred from is provided in Table A.21. MINING-D inferred 2 genes in most of the datasets, however in 3 datasets, 3 genes were inferred. The falsely inferred genes (shown in Table A.22) are substrings of TRBV genes and are inferred because the input sequences partially cover V genes.

Table A.20. Description of human TRB datasets. Datasets 1–6 belong to the "TCRB technical replicates of PBMC from four donors" project. Datasets 7–10 belong to the "Bone Marrow From Healthy Adults" project. Both projects are available at the immuneACCESS database by Adaptive Biotechnologies. For the "TCRB technical replicates of PBMC from four donors" project, we did not use datasets corresponding to the donor 4 because they are too small. The numbers of productive rearrangements and descriptions were taken from the immuneACCESS database.

#	Dataset	immuneACCESS sample name	# productive rearrangements	Description	# Inferred D genes
1	Subj1_1	Subject1_Tcells_aliquot24	117,292	Control, Sorted Cells, Subject 01, T cells, deep, gDNA, site 07	2
2	Subj1_2	Subject1_Tcells_aliquot26	132,807	Control, Sorted Cells, Subject 01, T cells, deep, gDNA, site 07	3
3	Subj2_1	Subject2_Tcells_aliquot24	112,172	Control, Sorted Cells, Subject 02, T cells, deep, gDNA, site 07	2
4	Subj2_2	Subject2_Tcells_aliquot26	130,789	Control, Sorted Cells, Subject 02, T cells, deep, gDNA, site 07	3
5	Subj3_1	Subject3_PBMC_aliquot24	83,347	Control, PBMC, Peripheral blood lymphocytes (PBL), Subject 03, T cells, deep, gDNA, site 07	2
6	Subj3_2	Subject3_PBMC_aliquot26	110,776	Control, PBMC, Peripheral blood lymphocytes (PBL), Subject 03, T cells, deep, gDNA, site 07	2
7	BM4385_1_TCRB	BM4385_1_TCRB	27,965	151-180 lbs, 18-35 Years, 5ft 11in - 6ft 2in, Bone marrow, HIV Neg, Hepatitis B Virus Negative, Hepatitis C Virus Negative, Hispanic, Hispanic Ethnicity, O	3
8	BM4359_1_TCRB	BM4359_1_TCRB	87,283	121-150 lbs, 18-35 Years, 5ft 6in - 5ft 10in, Bone marrow, HIV Neg, Hepatitis B Virus Negative, Hepatitis C Virus Negative, Hispanic, Hispanic Ethnicity	2
9	BM4359_1_TCRB	BM4359_1_TCRB	59,563	151-180 lbs, 18-35 Years, 5ft 6in - 5ft 10in, African American Ethnicity, African Race, B, Bone marrow, HIV Neg, Hepatitis B Virus Negative, Hepatitis C Virus Negative	2
10	BM4359_1_TCRB	BM4359_1_TCRB	36,703	18-35 Years, 181-210 lbs, 5ft 6in - 5ft 10in, A, Bone marrow, Caucasian, HIV Neg, Hepatitis B Virus Negative, Hepatitis C Virus Negative	2

Table A.21. Information about inferred D genes from TCR datasets using MINING-D.

Gene	Variant	Sequence	Datasets inferred in	Datasets NOT inferred in
TRBD1	TRBD1*01	GGGACAGGGGGC	–	BM4384
TRBD2	TRBD2*01	GGGACTAGCGGGGGGG	BM4374 BM4385	–
	TRBD2*02	GGGACTAGCGGGAGGG	–	BM4374 Bm4385

Table A.22. Information about falsely inferred D genes from TCR datasets using MINING-D. All three sequences are substrings of some V genes.

Sequence	Datasets inferred in	Comments
TGTATCTCTGTGCCACC	Subj2_2 BM4384	substring of TRBV23/OR9-2*01
TCTGTGCCAGCAGTTAC	Subj1_2	substring of TRBV6-2/TRBV6-3/TRBV6-5/TRBV6-6
TGTACTTCTGTGCCA	BM4385	substring of TRBV6-2/TRBV6-3/TRBV6-5/TRBV6-6

Bibliography

1. Cooper MD. The early history of B cells. Nat Rev Immunol. 2015;15(3):191-7. Epub 2015/02/06. doi: 10.1038/nri3801. PubMed PMID: 25656707.

2. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302(5909):575-81. doi: 10.1038/302575a0. PubMed PMID: 6300689.

3. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. Nat Protoc. 2016;11(9):1599-616. Epub 2016/08/04. doi: 10.1038/nprot.2016.093. PubMed PMID: 27490633.

4. Wang Y, Jackson KJ, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol Cell Biol. 2008;86(2):111-5. Epub 2007/11/27. doi: 10.1038/sj.icb.7100144. PubMed PMID: 18040280.

5. Ralph DK, Matsen FA. Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. 2017. doi: arXiv:1711.05843v2.

6. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. Nucleic Acids Res. 2012;40(17):e134. Epub 2012/05/27. doi: 10.1093/nar/gks457. PubMed PMID: 22641856; PubMed Central PMCID: PMCPMC3458526.

7. McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen FA. Quantifying evolutionary constraints on B-cell affinity maturation. Philos Trans R Soc Lond B Biol Sci. 2015;370(1676). doi: 10.1098/rstb.2014.0244. PubMed PMID: 26194758; PubMed Central PMCID: PMCPMC4528421.

8. Cui A, Di Niro R, Vander Heiden JA, Briggs AW, Adams K, Gilbert T, et al. A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. J Immunol. 2016;197(9):3566-74. Epub 2016/10/05. doi: 10.4049/jimmunol.1502263. PubMed PMID: 27707999; PubMed Central PMCID: PMCPMC5161250.

9. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012;13(5):363-73. Epub 2012/05/03. doi: 10.1038/gene.2012.12. PubMed PMID: 22551722.

10. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. Cell Host Microbe. 2013;13(6):691-700. doi: 10.1016/j.chom.2013.05.008. PubMed PMID: 23768493; PubMed Central PMCID: PMCPMC4136508.

11. Chang CJ, Chen CH, Chen BM, Su YC, Chen YT, Hershfield MS, et al. A genome-wide association study identifies a novel susceptibility locus for the immunogenicity of polyethylene glycol. Nat Commun. 2017;8(1):522. Epub 2017/09/12. doi: 10.1038/s41467-017-00622-4. PubMed PMID: 28900105; PubMed Central PMCID: PMCPMC5595925.

12. Boyd SD, Liu Y, Wang C, Martin V, Dunn-Walters DK. Human lymphocyte repertoires in ageing. Curr Opin Immunol. 2013;25(4):511-5. Epub 2013/08/28. doi: 10.1016/j.coi.2013.07.007. PubMed PMID: 23992996; PubMed Central PMCID: PMCPMC4811628.

13. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, et al. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. J Immunol. 2012;188(3):1333-40. Epub 2011/12/28. doi: 10.4049/jimmunol.1102097. PubMed PMID: 22205028; PubMed Central PMCID: PMCPMC4734744.

14. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci Rep. 2016;6:20842. Epub 2016/02/16. doi: 10.1038/srep20842. PubMed PMID: 26880249; PubMed Central PMCID: PMCPMC4754645.

15. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2009;37(Database issue):D1006-12. Epub 2008/10/31. doi: 10.1093/nar/gkn838. PubMed PMID: 18978023; PubMed Central PMCID: PMCPMC2686541.

16. Collins AM, Wang Y, Roskin KM, Marquis CP, Jackson KJ. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos Trans R Soc Lond B Biol Sci. 2015;370(1676). doi: 10.1098/rstb.2014.0236. PubMed PMID: 26194750; PubMed Central PMCID: PMCPMC4528413.

17. Muyldermans S, Smider VV. Distinct antibody species: structural differences creating therapeutic opportunities. Curr Opin Immunol. 2016;40:7-13. Epub 2016/02/27. doi: 10.1016/j.coi.2016.02.003. PubMed PMID: 26922135; PubMed Central PMCID: PMCPMC4884505.

18. de los Rios M, Criscitiello MF, Smider VV. Structural and genetic diversity in antibody repertoires from diverse species. Curr Opin Struct Biol. 2015;33:27-41. Epub 2015/07/17. doi: 10.1016/j.sbi.2015.06.002. PubMed PMID: 26188469.

19. Luo S, Yu JA, Li H, Song YS. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. Life Sci Alliance. 2019;2(2). Epub 2019/02/26. doi: 10.26508/lsa.201800221. PubMed PMID: 30808649; PubMed Central PMCID: PMCPMC6391684.

20. Yu Y, Ceredig R, Seoighe C. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. J Immunol. 2017;198(5):2202-10. Epub 2017/01/23. doi: 10.4049/jimmunol.1601710. PubMed PMID: 28115530.

21. Watson CT, Matsen FA, Jackson KJL, Bashir A, Smith ML, Glanville J, et al. Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data". J Immunol. 2017;198(9):3371-3. doi: 10.4049/jimmunol.1700306. PubMed PMID: 28416712.

22. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol. 2010;184(12):6986-92. Epub 2010/05/21. doi: 10.4049/jimmunol.1000445. PubMed PMID: 20495067; PubMed Central PMCID: PMCPMC4281569.

23. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc Natl Acad Sci U S A. 2015;112(8):E862-70. Epub 2015/02/09. doi: 10.1073/pnas.1417683112. PubMed PMID: 25675496; PubMed Central PMCID: PMCPMC4345584.

24. Corcoran MM, Phad GE, Vázquez Bernat, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun. 2016;7:13642. Epub 2016/12/20. doi: 10.1038/ncomms13642. PubMed PMID: 27995928; PubMed Central PMCID: PMCPMC5187446.

25. Zhang W, Wang IM, Wang C, Lin L, Chai X, Wu J, et al. IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data. Front Immunol. 2016;7:457. Epub 2016/11/04. doi: 10.3389/fimmu.2016.00457. PubMed PMID: 27867380; PubMed Central PMCID: PMCPMC5095119.

26. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, et al. Identification of Subject-Specific Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. Front Immunol. 2019;10:129. Epub 2019/02/13. doi: 10.3389/fimmu.2019.00129. PubMed PMID: 30814994; PubMed Central PMCID: PMCPMC6381938.

27. Khass M, Vale AM, Burrows PD, Schroeder HW. The sequences encoded by immunoglobulin diversity (D. Immunol Rev. 2018;284(1):106-19. doi: 10.1111/imr.12669. PubMed PMID: 29944758.

28. Safonova Y, Pevzner PA. Inference of Diversity Genes and Analysis of Non-canonical V(DD)J Recombination in Immunoglobulins. Front Immunol. 2019;10:987. Epub

2019/05/03. doi: 10.3389/fimmu.2019.00987. PubMed PMID: 31134072; PubMed Central PMCID: PMCPMC6516046.

29. Thörnqvist L, Ohlin M. Critical steps for computational inference of the 3'-end of novel alleles of immunoglobulin heavy chain variable genes - illustrated by an allele of IGHV3-7. Mol Immunol. 2018;103:1-6. Epub 2018/08/30. doi: 10.1016/j.molimm.2018.08.018. PubMed PMID: 30172112.

30. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, et al. Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. Front Immunol. 2019;10:435. Epub 2019/03/18. doi: 10.3389/fimmu.2019.00435. PubMed PMID: 30936866; PubMed Central PMCID: PMCPMC6431624.

31. Mitzenmacher M. A survey of results for deletion channels and related synchronization channels. Probability Surveys. 2009;6:1-33.

32. Levin M, Levander F, Palmason R, Greiff L, Ohlin M. Antibody-encoding repertoires of bone marrow and peripheral blood-a focus on IgE. J Allergy Clin Immunol. 2017;139(3):1026-30. Epub 2016/08/09. doi: 10.1016/j.jaci.2016.06.040. PubMed PMID: 27521279.

33. Ellebedy AH, Jackson KJ, Kissick HT, Nakaya HI, Davis CW, Roskin KM, et al. Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. Nat Immunol. 2016;17(10):1226-34. Epub 2016/08/15. doi: 10.1038/ni.3533. PubMed PMID: 27525369; PubMed Central PMCID: PMCPMC5054979.

34. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. J Immunol. 2017;198(6):2489-99. Epub 2017/02/08. doi: 10.4049/jimmunol.1601850. PubMed PMID: 28179494; PubMed Central PMCID: PMCPMC5340603.

35. Friedensohn S, Lindner JM, Cornacchione V, Iazeolla M, Miho E, Zingg A, et al. Synthetic Standards Combined With Error and Bias Correction Improve the Accuracy and Quantitative Resolution of Antibody Repertoire Sequencing in Human Naïve and Memory B Cells. Front Immunol. 2018;9:1401. Epub 2018/06/20. doi: 10.3389/fimmu.2018.01401. PubMed PMID: 29973938; PubMed Central PMCID: PMCPMC6019461.

36. Magri G, Comerma L, Pybus M, Sintes J, Lligé D, Segura-Garzón D, et al. Human Secretory IgM Emerges from Plasma Cells Clonally Related to Gut Memory B Cells and Targets Highly Diverse Commensals. Immunity. 2017;47(1):118-34.e8. Epub 2017/07/11. doi: 10.1016/j.immuni.2017.06.013. PubMed PMID: 28709802; PubMed Central PMCID: PMCPMC5519504.

37. Stern JN, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med. 2014;6(248):248ra107. doi: 10.1126/scitranslmed.3008879. PubMed PMID: 25100741; PubMed Central PMCID: PMCPMC4388137.

38. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, et al. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Rep. 2017;19(7):1467-78. doi: 10.1016/j.celrep.2017.04.054. PubMed PMID: 28514665.

39. Li X, Duan X, Yang K, Zhang W, Zhang C, Fu L, et al. Comparative Analysis of Immune Repertoires between Bactrian Camel's Conventional and Heavy-Chain Antibodies. PLoS One. 2016;11(9):e0161801. Epub 2016/09/02. doi: 10.1371/journal.pone.0161801. PubMed PMID: 27588755; PubMed Central PMCID: PMCPMC5010241.

40. VanDuijn MM, Dekker LJ, van IJcken WFJ, Sillevis Smitt PAE, Luider TM. Immune Repertoire after Immunization As Seen by Next-Generation Sequencing and Proteomics. Front Immunol. 2017;8:1286. Epub 2017/10/16. doi: 10.3389/fimmu.2017.01286. PubMed PMID: 29085363; PubMed Central PMCID: PMCPMC5650670.

41. Banerjee S, Shi H, Banasik M, Moon H, Lees W, Qin Y, et al. Evaluation of a novel multi-immunogen vaccine strategy for targeting 4E10/10E8 neutralizing epitopes on HIV-1 gp41 membrane proximal external region. Virology. 2017;505:113-26. Epub 2017/02/23. doi: 10.1016/j.virol.2017.02.015. PubMed PMID: 28237764; PubMed Central PMCID: PMCPMC5385849.

42. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, Pevzner PA. Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Reads. J Immunol. 2017;199(9):3369-80. Epub 2017/10/04. doi: 10.4049/jimmunol.1700485. PubMed PMID: 28978691; PubMed Central PMCID: PMCPMC5661950.

43. Kidd MJ, Jackson KJ, Boyd SD, Collins AM. DJ Pairing during VDJ Recombination Shows Positional Biases That Vary among Individuals with Differing IGHD Locus Immunogenotypes. J Immunol. 2016;196(3):1158-64. Epub 2015/12/23. doi: 10.4049/jimmunol.1501401. PubMed PMID: 26700767; PubMed Central PMCID: PMCPMC4724508.

44. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. Mol Immunol. 2017;87:12-22. Epub 2017/04/04. doi: 10.1016/j.molimm.2017.03.012. PubMed PMID: 28388445.

45. Mattson MP, Allison DB, Fontana L, Harvie M, Longo VD, Malaisse WJ, et al. Meal frequency and timing in health and disease. Proc Natl Acad Sci U S A. 2014;111(47):16647-53. Epub 2014/11/19. doi: 10.1073/pnas.1413965111. PubMed PMID: 25404320; PubMed Central PMCID: PMCPMC4250148.

46. Di Francesco A, Di Germanio C, Bernier M, de Cabo R. A time to fast. Science. 2018;362(6416):770-5. Epub 2018/11/18. doi: 10.1126/science.aau2095. PubMed PMID: 30442801.

47. de Cabo R, Mattson MP. Effects of Intermittent Fasting on Health, Aging, and Disease. N Engl J Med. 2019;381(26):2541-51. Epub 2019/12/28. doi: 10.1056/NEJMra1905136. PubMed PMID: 31881139.

48. Walford RL, Mock D, Verdery R, MacCallum T. Calorie restriction in biosphere 2: alterations in physiologic, hematologic, hormonal, and biochemical parameters in humans restricted for a 2-year period. J Gerontol A Biol Sci Med Sci. 2002;57(6):B211-24. Epub 2002/05/23. doi: 10.1093/gerona/57.6.b211. PubMed PMID: 12023257.

49. Fontana L, Meyer TE, Klein S, Holloszy JO. Long-term calorie restriction is highly effective in reducing the risk for atherosclerosis in humans. Proc Natl Acad Sci U S A. 2004;101(17):6659-63. Epub 2004/04/21. doi: 10.1073/pnas.0308291101. PubMed PMID: 15096581; PubMed Central PMCID: PMCPMC404101.

50. Rynders CA, Thomas EA, Zaman A, Pan Z, Catenacci VA, Melanson EL. Effectiveness of Intermittent Fasting and Time-Restricted Feeding Compared to Continuous Energy Restriction for Weight Loss. Nutrients. 2019;11(10). Epub 2019/10/17. doi: 10.3390/nu11102442. PubMed PMID: 31614992; PubMed Central PMCID: PMCPMC6836017.

51. Welton S, Minty R, O'Driscoll T, Willms H, Poirier D, Madden S, et al. Intermittent fasting and weight loss: Systematic review. Can Fam Physician. 2020;66(2):117-25. Epub 2020/02/16. PubMed PMID: 32060194; PubMed Central PMCID: PMCPMC7021351.

52. Goldhamer AC, Lisle DJ, Sultana P, Anderson SV, Parpia B, Hughes B, et al. Medically supervised water-only fasting in the treatment of borderline hypertension. J Altern Complement Med. 2002;8(5):643-50. Epub 2002/12/10. doi: 10.1089/107555302320825165. PubMed PMID: 12470446.

53. Muller H, de Toledo FW, Resch KL. Fasting followed by vegetarian diet in patients with rheumatoid arthritis: a systematic review. Scand J Rheumatol. 2001;30(1):1-10. Epub 2001/03/17. doi: 10.1080/030097401750065256. PubMed PMID: 11252685.

54. Kjeldsen-Kragh J, Haugen M, Borchgrevink CF, Laerum E, Eek M, Mowinkel P, et al. Controlled trial of fasting and one-year vegetarian diet in rheumatoid arthritis. Lancet. 1991;338(8772):899-902. Epub 1991/10/12. doi: 10.1016/0140-6736(91)91770-u. PubMed PMID: 1681264.

55. Rochon J, Bales CW, Ravussin E, Redman LM, Holloszy JO, Racette SB, et al. Design and conduct of the CALERIE study: comprehensive assessment of the long-term effects of reducing intake of energy. J Gerontol A Biol Sci Med Sci. 2011;66(1):97-108. Epub 2010/10/07. doi: 10.1093/gerona/glq168. PubMed PMID: 20923909; PubMed Central PMCID: PMCPMC3032519.

56. Martin CK, Bhapkar M, Pittas AG, Pieper CF, Das SK, Williamson DA, et al. Effect of Calorie Restriction on Mood, Quality of Life, Sleep, and Sexual Function in Healthy Nonobese Adults: The CALERIE 2 Randomized Clinical Trial. JAMA Intern Med. 2016;176(6):743-52. Epub 2016/05/03. doi: 10.1001/jamainternmed.2016.1189. PubMed PMID: 27136347; PubMed Central PMCID: PMCPMC4905696.

57. Das SK, Roberts SB, Bhapkar MV, Villareal DT, Fontana L, Martin CK, et al. Body-composition changes in the Comprehensive Assessment of Long-term Effects of Reducing Intake of Energy (CALERIE)-2 study: a 2-y randomized controlled trial of calorie restriction in nonobese humans. Am J Clin Nutr. 2017;105(4):913-27. Epub 2017/02/24. doi:

10.3945/ajcn.116.137232. PubMed PMID: 28228420; PubMed Central PMCID: PMCPMC5366044.

58. Most J, Gilmore LA, Smith SR, Han H, Ravussin E, Redman LM. Significant improvement in cardiometabolic health in healthy nonobese individuals during caloric restriction-induced weight loss and weight loss maintenance. Am J Physiol Endocrinol Metab. 2018;314(4):E396-E405. Epub 2018/01/20. doi: 10.1152/ajpendo.00261.2017. PubMed PMID: 29351490; PubMed Central PMCID: PMCPMC5966756.

59. Kraus WE, Bhapkar M, Huffman KM, Pieper CF, Krupa Das S, Redman LM, et al. 2 years of calorie restriction and cardiometabolic risk (CALERIE): exploratory outcomes of a multicentre, phase 2, randomised controlled trial. Lancet Diabetes Endocrinol. 2019;7(9):673-83. Epub 2019/07/16. doi: 10.1016/S2213-8587(19)30151-2. PubMed PMID: 31303390; PubMed Central PMCID: PMCPMC6707879.

60. Longo VD, Mattson MP. Fasting: molecular mechanisms and clinical applications. Cell Metab. 2014;19(2):181-92. Epub 2014/01/21. doi: 10.1016/j.cmet.2013.12.008. PubMed PMID: 24440038; PubMed Central PMCID: PMCPMC3946160.

61. Yore MM, Syed I, Moraes-Vieira PM, Zhang T, Herman MA, Homan EA, et al. Discovery of a class of endogenous mammalian lipids with anti-diabetic and anti-inflammatory effects. Cell. 2014;159(2):318-32. doi: 10.1016/j.cell.2014.09.035. PubMed PMID: 25303528; PubMed Central PMCID: PMCPMC4260972.

62. Boden G. Obesity, insulin resistance and free fatty acids. Curr Opin Endocrinol Diabetes Obes. 2011;18(2):139-43. doi: 10.1097/MED.0b013e3283444b09. PubMed PMID: 21297467; PubMed Central PMCID: PMCPMC3169796.

63. Roden M, Price TB, Perseghin G, Petersen KF, Rothman DL, Cline GW, et al. Mechanism of free fatty acid-induced insulin resistance in humans. J Clin Invest. 1996;97(12):2859-65. doi: 10.1172/JCI118742. PubMed PMID: 8675698; PubMed Central PMCID: PMCPMC507380.

64. Boden G. Effects of free fatty acids (FFA) on glucose metabolism: significance for insulin resistance and type 2 diabetes. Exp Clin Endocrinol Diabetes. 2003;111(3):121-4. doi: 10.1055/s-2003-39781. PubMed PMID: 12784183.

65. Trost Z, Vangronsveld K, Linton SJ, Quartana PJ, Sullivan MJ. Cognitive dimensions of anger in chronic pain. Pain. 2012;153(3):515-7. Epub 2011/12/02. doi: 10.1016/j.pain.2011.10.023. PubMed PMID: 22136750.

66. Boylan JM, Ryff CD. Varieties of anger and the inverse link between education and inflammation: toward an integrative framework. Psychosom Med. 2013;75(6):566-74. Epub 2013/06/13. doi: 10.1097/PSY.0b013e31829683bd. PubMed PMID: 23766379; PubMed Central PMCID: PMCPMC3702653.

67. Fernandez E. The relationship between anger and pain. Curr Pain Headache Rep. 2005;9(2):101-5. PubMed PMID: 15745619.

68. Albanese E, Matthews KA, Zhang J, Jacobs DR, Whitmer RA, Wadley VG, et al. Hostile attitudes and effortful coping in young adulthood predict cognition 25 years later. Neurology.

2016;86(13):1227-34. Epub 2016/03/02. doi: 10.1212/WNL.0000000000002517. PubMed PMID: 26935891; PubMed Central PMCID: PMCPMC4818565.

69. Tsuchiyama K, Terao T, Wang Y, Hoaki N, Goto S. Relationship between hostility and subjective sleep quality. Psychiatry Res. 2013;209(3):545-8. Epub 2013/04/09. doi: 10.1016/j.psychres.2013.03.019. PubMed PMID: 23582207.

70. Barefoot JC, Larsen S, von der Lieth L, Schroll M. Hostility, incidence of acute myocardial infarction, and mortality in a sample of older Danish men and women. Am J Epidemiol. 1995;142(5):477-84. PubMed PMID: 7677126.

71. Everson SA, Kauhanen J, Kaplan GA, Goldberg DE, Julkunen J, Tuomilehto J, et al. Hostility and increased risk of mortality and acute myocardial infarction: the mediating role of behavioral risk factors. Am J Epidemiol. 1997;146(2):142-52. PubMed PMID: 9230776.

72. Forbes D, Creamer M, Hawthorne G, Allen N, McHugh T. Comorbidity as a predictor of symptom change after treatment in combat-related posttraumatic stress disorder. J Nerv Ment Dis. 2003;191(2):93-9. doi: 10.1097/01.NMD.0000051903.60517.98. PubMed PMID: 12586962.

73. Teten AL, Miller LA, Stanford MS, Petersen NJ, Bailey SD, Collins RL, et al. Characterizing aggression and its association to anger and hostility among male veterans with post-traumatic stress disorder. Mil Med. 2010;175(6):405-10. PubMed PMID: 20572472.

74. Spielberger CD, Jacobs G, Russell S, Crane RS. Assessment of anger: The State-Trait Anger Scale . Hillsdale, NJ : Lawrence Erlbaum: Advances in personality assessment 1983. p. 159 - 87.

75. Orth U, Wieland E. Anger, hostility, and posttraumatic stress disorder in trauma-exposed adults: a meta-analysis. J Consult Clin Psychol. 2006;74(4):698-706. doi: 10.1037/0022-006X.74.4.698. PubMed PMID: 16881777.

76. MacManus D, Rona R, Dickson H, Somaini G, Fear N, Wessely S. Aggressive and violent behavior among military personnel deployed to Iraq and Afghanistan: prevalence and link with deployment and combat exposure. Epidemiol Rev. 2015;37:196-212. Epub 2015/01/22. doi: 10.1093/epirev/mxu006. PubMed PMID: 25613552.

77. Dennis PA, Dennis NM, Van Voorhees EE, Calhoun PS, Dennis MF, Beckham JC. Moral transgression during the Vietnam War: a path analysis of the psychological impact of veterans' involvement in wartime atrocities. Anxiety Stress Coping. 2017;30(2):188-201. Epub 2016/09/19. doi: 10.1080/10615806.2016.1230669. PubMed PMID: 27580161; PubMed Central PMCID: PMCPMC5299042.

78. Wilk JE, Quartana PJ, Clarke-Walper K, Kok BC, Riviere LA. Aggression in US soldiers post-deployment: Associations with combat exposure and PTSD and the moderating role of trait anger. Aggress Behav. 2015;41(6):556-65. Epub 2015/07/23. doi: 10.1002/ab.21595. PubMed PMID: 26205643.

79. Angkaw AC, Ross BS, Pittman JO, Kelada AM, Valencerina MA, Baker DG. Post-traumatic stress disorder, depression, and aggression in OEF/OIF veterans. Mil Med. 2013;178(10):1044-50. doi: 10.7205/MILMED-D-13-00061. PubMed PMID: 24083916.

80. Norman SB, Schmied E, Larson GE. Physical aggression among post-9/11 veterans. Military behavioral health. 2015;3(1):47-54.

81. Weiss NH, Connolly KM, Gratz KL, Tull MT. The Role of Impulsivity Dimensions in the Relation Between Probable Posttraumatic Stress Disorder and Aggressive Behavior Among Substance Users. J Dual Diagn. 2017;13(2):109-18. Epub 2017/02/08. doi: 10.1080/15504263.2017.1293310. PubMed PMID: 28368772; PubMed Central PMCID: PMCPMC5472350.

82. Elbogen EB, Johnson SC, Newton VM, Fuller S, Wagner HR, Beckham JC, et al. Self-report and longitudinal predictors of violence in Iraq and Afghanistan war era veterans. J Nerv Ment Dis. 2013;201(10):872-6. doi: 10.1097/NMD.0b013e3182a6e76b. PubMed PMID: 24080674; PubMed Central PMCID: PMCPMC3919673.

83. Teten AL, Miller LA, Bailey SD, Dunn NJ, Kent TA. Empathic deficits and alexithymia in trauma-related impulsive aggression. Behav Sci Law. 2008;26(6):823-32. doi: 10.1002/bsl.843. PubMed PMID: 19039794.

84. Taft CT, Weatherill RP, Woodward HE, Pinto LA, Watkins LE, Miller MW, et al. Intimate partner and general aggression perpetration among combat veterans presenting to a posttraumatic stress disorder clinic. Am J Orthopsychiatry. 2009;79(4):461-8. doi: 10.1037/a0016657. PubMed PMID: 20099937; PubMed Central PMCID: PMCPMC3561901.

85. Dursa EK, Reinhard MJ, Barth SK, Schneiderman AI. Prevalence of a positive screen for PTSD among OEF/OIF and OEF/OIF-era veterans in a large population-based cohort. J Trauma Stress. 2014;27(5):542-9. Epub 2014/09/29. doi: 10.1002/jts.21956. PubMed PMID: 25267288.

86. Lapierre CB, Schwegler AF, Labauve BJ. Posttraumatic stress and depression symptoms in soldiers returning from combat operations in Iraq and Afghanistan. J Trauma Stress. 2007;20(6):933-43. doi: 10.1002/jts.20278. PubMed PMID: 18157882.

87. Jakupcak M, Conybeare D, Phelps L, Hunt S, Holmes HA, Felker B, et al. Anger, hostility, and aggression among Iraq and Afghanistan War veterans reporting PTSD and subthreshold PTSD. J Trauma Stress. 2007;20(6):945-54. doi: 10.1002/jts.20258. PubMed PMID: 18157891.

88. Kwan J, Jones M, Somaini G, Hull L, Wessely S, Fear NT, et al. Post-deployment family violence among UK military personnel. Psychol Med. 2017:1-11. Epub 2017/12/19. doi: 10.1017/S0033291717003695. PubMed PMID: 29254510.

89. Taft CT, Kaloupek DG, Schumm JA, Marshall AD, Panuzio J, King DW, et al. Posttraumatic stress disorder symptoms, physiological reactivity, alcohol problems, and aggression among military veterans. J Abnorm Psychol. 2007;116(3):498-507. doi: 10.1037/0021-843X.116.3.498. PubMed PMID: 17696706.

90. Watkins LE, Sippel LM, Pietrzak RH, Hoff R, Harpaz-Rotem I. Co-occurring aggression and suicide attempt among veterans entering residential treatment for PTSD: The role of PTSD symptom clusters and alcohol misuse. J Psychiatr Res. 2017;87:8-14. Epub 2016/12/09. doi: 10.1016/j.jpsychires.2016.12.009. PubMed PMID: 27984702.

91. Stappenbeck CA, Hellmuth JC, Simpson T, Jakupcak M. The Effects of Alcohol Problems, PTSD, and Combat Exposure on Nonphysical and Physical Aggression Among Iraq and Afghanistan War Veterans. Psychol Trauma. 2014;6(1):65-72. doi: 10.1037/a0031468. PubMed PMID: 25225593; PubMed Central PMCID: PMCPMC4163149.

92. McCubbin JA, Zinzow HM, Hibdon MA, Nathan AW, Morrison AV, Hayden GW, et al. Subclinical Posttraumatic Stress Disorder Symptoms: Relationships with Blood Pressure, Hostility, and Sleep. Cardiovasc Psychiatry Neurol. 2016;2016:4720941. Epub 2016/06/15. doi: 10.1155/2016/4720941. PubMed PMID: 27403340; PubMed Central PMCID: PMCPMC4925987.

93. Van Voorhees EE, Dennis PA, Neal LC, Hicks TA, Calhoun PS, Beckham JC, et al. Posttraumatic Stress Disorder, Hostile Cognitions, and Aggression in Iraq/Afghanistan Era Veterans. Psychiatry. 2016;79(1):70-84. doi: 10.1080/00332747.2015.1123593. PubMed PMID: 27187514; PubMed Central PMCID: PMCPMC4973515.

94. Vrana SR, Hughes JW, Dennis MF, Calhoun PS, Beckham JC. Effects of posttraumatic stress disorder status and covert hostility on cardiovascular responses to relived anger in women with and without PTSD. Biol Psychol. 2009;82(3):274-80. Epub 2009/08/27. doi: 10.1016/j.biopsycho.2009.08.008. PubMed PMID: 19716397; PubMed Central PMCID: PMCPMC2767439.

95. Nabi H, Singh-Manoux A, Ferrie JE, Marmot MG, Melchior M, Kivimäki M. Hostility and depressive mood: results from the Whitehall II prospective cohort study. Psychol Med. 2010;40(3):405-13. Epub 2009/07/17. doi: 10.1017/S0033291709990432. PubMed PMID: 19607752; PubMed Central PMCID: PMCPMC2749841.

96. Olatunji BO, Ciesielski BG, Tolin DF. Fear and loathing: a meta-analytic review of the specificity of anger in PTSD. Behav Ther. 2010;41(1):93-105. Epub 2009/09/02. doi: 10.1016/j.beth.2009.01.004. PubMed PMID: 20171331.

97. Taft CT, Creech SK, Murphy CM. Anger and aggression in PTSD. Curr Opin Psychol. 2017;14:67-71. Epub 2016/12/01. doi: 10.1016/j.copsyc.2016.11.008. PubMed PMID: 28813322.

98. Spielberger CD. STAXI-2 State-Trait Anger Expression Inventory-2: Professional Manual. Lutz, FL: Psychological Assessment Resources, Inc.; 1999.

99. Kulkarni M, Porter KE, Rauch SA. Anger, dissociation, and PTSD among male veterans entering into PTSD treatment. J Anxiety Disord. 2012;26(2):271-8. Epub 2011/12/16. doi: 10.1016/j.janxdis.2011.12.005. PubMed PMID: 22245698.

100. Meffert SM, Metzler TJ, Henn-Haase C, McCaslin S, Inslicht S, Chemtob C, et al. A prospective study of trait anger and PTSD symptoms in police. J Trauma Stress. 2008;21(4):410-6. doi: 10.1002/jts.20350. PubMed PMID: 18720397; PubMed Central PMCID: PMCPMC3974928.

101. Birkley EL, Schumm JA. Posttraumatic Stress Disorder, Aggressive Behavior, and Anger: Recent Findings and Treatment Recommendations. Current Treatment Options in Psychiatry. 2016;3(1):48-59.

102. Hellmuth JC, Stappenbeck CA, Hoerster KD, Jakupcak M. Modeling PTSD symptom clusters, alcohol misuse, anger, and depression as they relate to aggression and suicidality in returning U.S. veterans. J Trauma Stress. 2012;25(5):527-34. doi: 10.1002/jts.21732. PubMed PMID: 23073972; PubMed Central PMCID: PMCPMC4068010.

103. Ethical principles of psychologists and code of conduct. Am Psychol. 2002;57(12):1060-73. PubMed PMID: 12613157.

104. Blake DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS, et al. The development of a Clinician-Administered PTSD Scale. J Trauma Stress. 1995;8(1):75-90. PubMed PMID: 7712061.

105. Mueser KT, Salyers MP, Rosenberg SD, Ford JD, Fox L, Carty P. Psychometric evaluation of trauma and posttraumatic stress disorder assessments in persons with severe mental illness. Psychol Assess. 2001;13(1):110-7. PubMed PMID: 11281032.

106. Weathers FW, Keane TM, Davidson JR. Clinician-administered PTSD scale: a review of the first ten years of research. Depress Anxiety. 2001;13(3):132-56. PubMed PMID: 11387733.

107. Beck A, Steer R, Brown G. Manual for the Beck Depression Inventory–II. San Antonio, TX: Psychological Corporation; 1996.

108. Cook WW, Medley DM. Proposed hostility and Pharisaic-virtue scales for the MMPI. US: American Psychological Association; 1954. p. 414-8.

109. Barefoot JC, Dodge KA, Peterson BL, Dahlstrom WG, Williams RB. The Cook-Medley hostility scale: item content and ability to predict survival. Psychosom Med. 1989;51(1):46-57. PubMed PMID: 2928460.

110. Sorgi P, Ratey J, Knoedler DW, Markert RJ, Reichman M. Rating aggression in the clinical setting. A retrospective adaptation of the Overt Aggression Scale: preliminary results. J Neuropsychiatry Clin Neurosci. 1991;3(2):S52-6. PubMed PMID: 1687961.

111. Yudofsky SC, Silver JM, Jackson W, Endicott J, Williams D. The Overt Aggression Scale for the objective rating of verbal and physical aggression. Am J Psychiatry. 1986;143(1):35-9. doi: 10.1176/ajp.143.1.35. PubMed PMID: 3942284.

112. Goldberg BR, Serper MR, Sheets M, Beech D, Dill C, Duffy KG. Predictors of aggression on the psychiatric inpatient service: self-esteem, narcissism, and theory of mind deficits. J Nerv Ment Dis. 2007;195(5):436-42. doi: 10.1097/01.nmd.0000253748.47641.99. PubMed PMID: 17502810.

113. Serper M, Beech DR, Harvey PD, Dill C. Neuropsychological and symptom predictors of aggression on the psychiatric inpatient service. J Clin Exp Neuropsychol. 2008;30(6):700-9. Epub 2008/01/22. doi: 10.1080/13803390701684554. PubMed PMID: 18608673.

114. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods. 2008;40(3):879-91. PubMed PMID: 18697684.

115. Resick PA, Nishith P, Weaver TL, Astin MC, Feuer CA. A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. J Consult Clin Psychol. 2002;70(4):867-79. PubMed PMID: 12182270; PubMed Central PMCID: PMCPMC2977927.

116. US Department of Veterans Affairs. Available from: https://www.va.gov/vetdata/.

117. Gusfield D. Algorithms on strings, trees, and sequences : computer science and computational biology. Cambridge England ; New York: Cambridge University Press; 1997. xviii, 534 pages p.

118. Souto-Carneiro MM, Sims GP, Girschik H, Lee J, Lipsky PE. Developmental changes in the human heavy chain CDR3. J Immunol. 2005;175(11):7425-36. doi: 10.4049/jimmunol.175.11.7425. PubMed PMID: 16301650.

119. Briney BS, Willis JR, Hicar MD, Thomas JW, Crowe JE. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. Immunology. 2012;137(1):56-64. doi: 10.1111/j.1365-2567.2012.03605.x. PubMed PMID: 22612413; PubMed Central PMCID: PMCPMC3449247.

120. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. Philos Trans R Soc Lond B Biol Sci. 2015;370(1676). doi: 10.1098/rstb.2014.0243. PubMed PMID: 26194757; PubMed Central PMCID: PMCPMC4528420.

121. Safonova Y, Lapidus A, Lill J. IgSimulator: a versatile immunosequencing simulator. Bioinformatics. 2015;31(19):3213-5. Epub 2015/05/25. doi: 10.1093/bioinformatics/btv326. PubMed PMID: 26007226.