

Automated analysis of immunosequencing datasets reveals novel immunoglobulin D genes across diverse species

Vinnu Bhardwaj¹, Yana Safonova^{2,3}, Massimo Franceschetti¹, Ramesh Rao¹, and Pavel A. Pevzner^{3,*}

¹ Electrical and Computer Engineering Department, University of California San Diego, San Diego, USA

² Center for Information Theory and Applications, University of California San Diego, San Diego, USA

³ Computer Science and Engineering Department, University of California San Diego, San Diego, USA

* Corresponding author: ppevzner@ucsd.edu

Abstract

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. Although reconstruction of V and J genes is a well-studied problem, the more challenging task of reconstructing D genes remained open until the IgScout algorithm was developed in 2019. In this work, we address limitations of IgScout by developing a probabilistic MINING-D algorithm for D gene reconstruction, apply it to hundreds of immunosequencing datasets from multiple species, and validate the newly inferred D genes by analyzing diverse whole genome sequencing datasets.

Author summary

Antibodies 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. Most of the previous immunogenomics studies rely on the reference germline genes in the immunoglobulin locus rather than the germline genes in a *specific patient*. This approach is deficient since the set of known germline genes is incomplete (particularly for non-Europeans humans and non-human species) and contains alleles that resulted from sequencing and annotation errors. The problem of *de novo* inference of diversity (D) genes from immunosequencing data remained open until the IgScout algorithm was developed in 2019. We address limitations of IgScout by developing a probabilistic MINING-D algorithm for D gene reconstruction and infer multiple D genes across multiple species that are not present in standard databases.

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 has emerged as the method of choice for generating millions of reads that sample antibody repertoires and provides insights into monitoring immune response to disease and vaccination [3]. Information about all germline genes in an individual is a pre-requisite for analyzing immunogenomics data. However, nearly all previous immunogenomics studies relied on the population-level germline genes rather than the germline genes in a specific individual from which the immunosequencing data originated. This approach is deficient, since the set of known germline genes

is incomplete (particularly for non-Europeans) and contains alleles that resulted from sequencing and annotation errors [4, 5]. Moreover, it is non-trivial 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]. These errors hide the identity of the individual germline genes, make it difficult to analyze *somatic hypermutations (SHM)*, and complicate studies of antibody evolution [6-8].

Personalized immunogenomics (i.e., identifying individual germline genes) 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]. However, since the International ImMunoGeneTics (IMGT) database [15] is incomplete even in the case of well-studied human germline genes [16], there still exist unknown human allelic variants that are difficult to differentiate from SHMs. In the case of immunologically important but less studied model organisms [17, 18], the germline genes remain largely unknown. Unfortunately, since assembling the highly repetitive immunoglobulin locus from whole genome sequencing data faces challenges [19], efforts such as the 1000 Genomes Project have resulted in only limited progress towards inferring the population-wide census of germline 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 allelic 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 key roles in B cell development, antigen binding site diversity, and antibody production [27]. Although Safanova and Pevzner [28] recently developed the IgScout algorithm for *de novo* inference of D genes using immunosequencing data, 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 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.

Methods

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 $|\cdot|$ denotes the length of a string) such that the first p and the last q symbols of s are removed (Figure 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.

D gene inference and the trace reconstruction problem. Given a set $C = \{c_1, c_2, c_3, \dots, c_N\}$ of independently generated instances of the modified strings generated from an unknown set of seed strings $S = \{s_1, s_2, \dots, s_M\}$, 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 [30] *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.

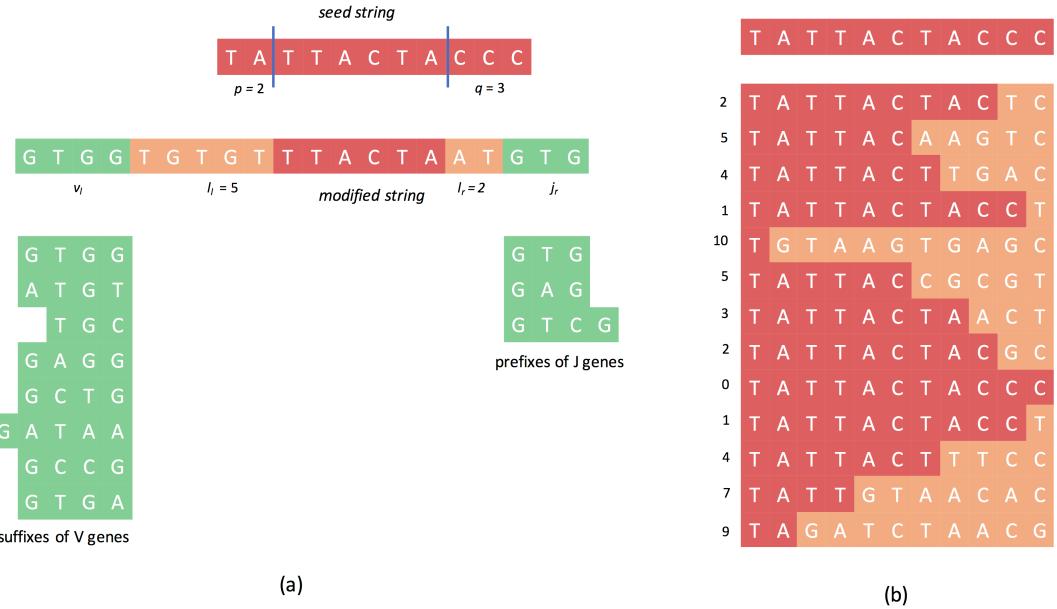


Figure 1. Transformation of a seed string representing a D gene into a modified string representing a CDR3 (a) and a set of modified strings generated according to a simple probabilistic model (b). (a) The symbols in red, yellow, and green in the modified string denote the symbols from the truncated 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.”

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

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

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

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

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

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

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

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:

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

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 1(b). Given a string s , $\text{score}(C|s)$ can be computed in $O(|s|*N)$ time.

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 $\text{sim}_t(C, s)$. Given a string s ,

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

We use this recurrence to efficiently compute $\text{score}(C|s)$ for each string s from C using dynamic programming. We construct a *trie* of all strings in C [31]. Each vertex in the trie is a t -prefix s^t of a string from C , and we recursively compute $\text{score}(C^t|s^t)$ 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 2). All scores can be computed by a single Depth First Search, assuming that all values $\text{sim}_i(C, s)$ are computed during the construction of the trie.

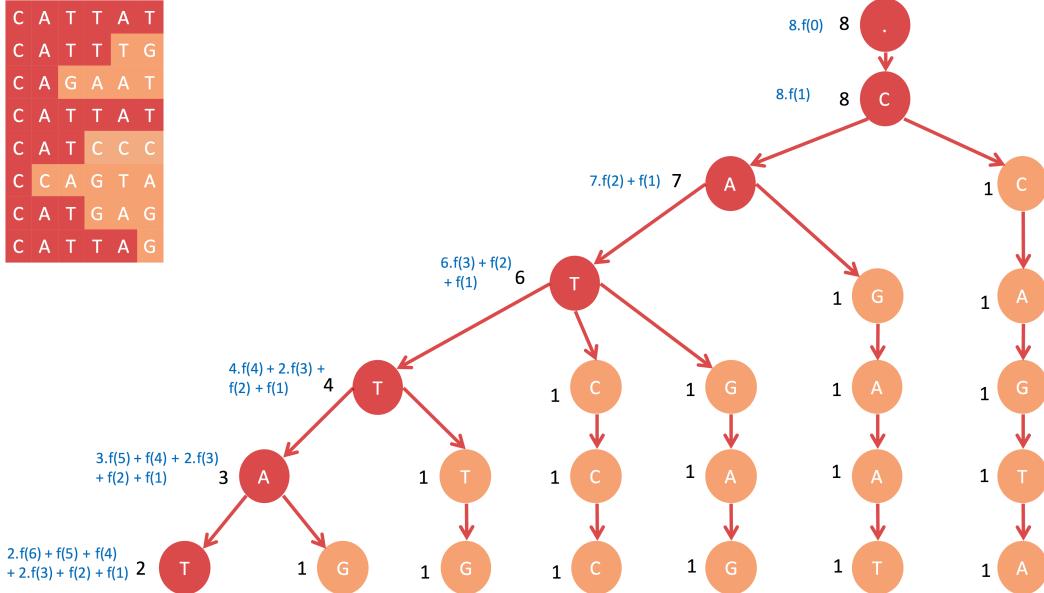


Figure 2. Illustration of the algorithm for solving the String Reconstruction Problem. The set of modified strings is shown on the left, and their trie is shown on the right. The string associated with each vertex is the one that is formed by traversing from the root node to the vertex. The number of 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.

Greedy algorithm for D gene inference. Although the objective function in (6) can be efficiently maximized by the above algorithm, it is unclear how to generalize this 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 **I**Nference of **I**mmuNoglobulin **G**enes - **D**) inspired by the above greedy algorithm and considering the complexities of real CDR3s.

MINING-D algorithm. Figure 3 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 3). 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.

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 3).

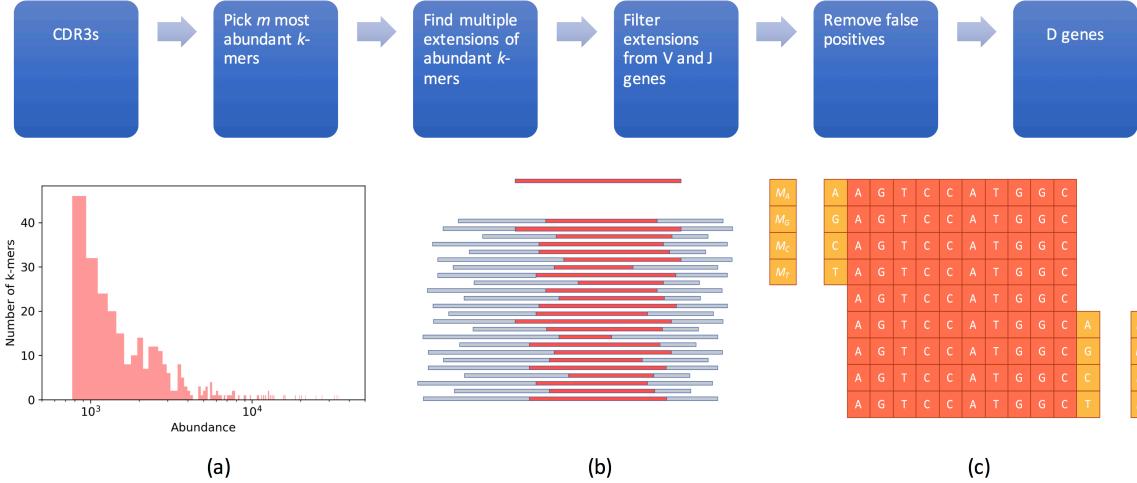


Figure 3. Outline of the MINING-D algorithm. (Top) MINING-D pipeline. (Bottom, a) Abundances of the 300 most abundant 10-mers in the MOUSE dataset vary from 770 to 34451. (Bottom, b) A D gene (top) and its truncated substrings in various CDR3s of varying lengths (bottom). The red part of a CDR3 is the “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.

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 3). 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.

Finding multiple extensions of seed k -mers. Some highly abundant k -mers can be substrings of multiple D genes as shown in Figure 4. 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 4. The 10-mer ATTACTACGG is present in two human D gene segments.

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 5. 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.

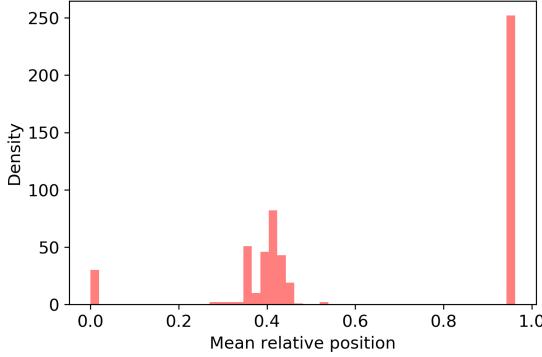


Figure 5. The mean relative positions of the extensions in the MOUSE dataset. The relative positions of the extensions form three clusters each corresponding to one of the V, D, and J gene segments.

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 6 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).

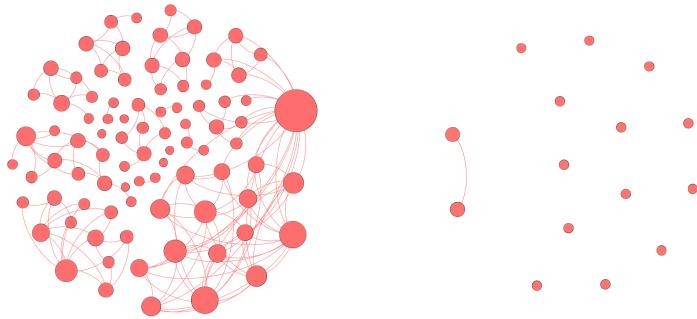


Figure 6. Similarity graph on all extensions corresponding to D genes before filtering extensions or clique merging for the MOUSE dataset (left) and after filtering extensions and merging cliques (right). The size of a node represents its degree.

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.

Results

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 1.

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	

Table 1. Meta-categories of datasets.

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.

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.

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

Table 2. Information about the D genes in the IMGT database for various species.

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 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 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 (see Supplemental Note: IgScout vs MINING-D). IgScout failed to reconstruct seven D genes from the IMGT database from all the datasets, whereas MINING-D only missed three genes.

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)

Table 3. Information about inferred D genes. The number of novel genes and variations validated using genomic data (procedure described later) are shown.

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 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~~TTTGGGGAGTTATCGTTAT was inferred as a novel variation of the gene IGHD3-16*02 (N_Var (IGHD3-16*02)-0) from immunosequencing data, the full sequence **GT**~~A~~TATTGATTAC~~A~~TTTGGGGAGTTATCGTTAT**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 TTTGGAGTGTTAT
N_Gene-1	ATTACAAT TTTGGACTGGTTATTAT

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

N_Gene-0	GTATTACAAT TTTGGAGTGTTATTACACC
N_Gene-1	GTATTACAAT TTTGGACTGGTTATTATACC .

Human	
IGHD3-10*01 Original GTATTACTATGGTCGGGAGTTATTATAAC N_Var-3 GTATTACTATGGTC AGGGAGTTATTATAAC	IGHD3-16*02 Original GTATTATGATTACGTTGGGGAGTTATCGTTATACC N_Var-0 ---TTATGATTAC ATTGGGGAGTTATCGTTAT ---
Camel	
IGHD3*01 (Alpaca) Original GTATTACTACTGCTCAGGCTATGGTGTATGAC N_Var-1 ---- GACTGCTATCAGGCTCTGGTGTATG -- N_Var-0 --- TGACTACTGTCAGGCTCTGGTGT -----	IGHD2*01 (Alpaca) Original ACATACTATAGGGTAGTTACTACACC N_Var-1 --ATAT TGTAGTGGTGGTTACTGCTAC -- N_Var-0 -CATACTATAGGGT GGTTACTAC -----
IGHD4*01 (Alpaca) Original TTACTATAGCGACTATGAC N_Var-1 C TACTATAGCGACTATG-- N_Var-0 -TACTATA ACGAATATG --	IGHD6*01 (Alpaca) Original GTACGGTAGTAGCTGGTAC N_Var-2 GTACGGT GGTAGCTGGTAC
IGHD5*01 (Alpaca) Original AGACTACGGGTTGGGTAC N_Var-0 ---- TATGGGTT-GGGTAC	
Rhesus Macaque	
IGHD1S39*01 Original GGTATAGGGAACTACAAC N_Var-0 ----- AGTGGGAGCTAC ---	IGHD3S18*01 Original GTACTGGGTGATTATTATGAC N_Var-0 -- ACTGGAGTGATTATTA ----
IGHD5S3*01 Original GTGGATACAGTGGGTACAGTTAC N_Var-0 --- GATACAGCGGGTACAGT ---	IGHD2S11*01 Original AGAATATTGTAGTAGTACTACTGCTCCTCC N_Var-0 ---- ATTGTAGTGGTACTACTGCT ----
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 TTTATTACTAC CGTAGCTACg	
Rat	
IGHD1-3*01 Original TTTTTAACTATGGTAGCTAC N_Var-0 -TTTTAACTAC CGGTAGCTAC	IGHD1-9*01 Original TACATACTATGGGTATAACTAC- N_Var-1 --CATACTAC GGGTATA CCTACg
IGHD1-12*02 Original TTTATTACTATGATGGTAGTTATTACTAC- N_Var-0 -TTATTACTATGATGGTAC TTTATTACTACg	
Rabbit	
IGHD6-1*01 Original ----- GTTACTATAGTTATGGTTATGCTTATGCTAC N_Var-4 TTACTATACTATGGTTATGCTGGTTATGCTTATGCTAC ----- N_Var-3 TTA-----TGCTGGTTATGCTGGTTATGCTAC -----	IGHD1-1*01 Original GCATATACTAGTAGTAGTGGTTATTATATAC N_Var-2 GCATAT GCTAGTAGTAGTGGTTATTAT ----

Table 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 TTTATTACTACGATGGTAGCTACg is only present as TTTATTACTACGATGGTAGCTAC in the genomic reads.

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. Since paired genomic and immunosequencing datasets were not available, genomic and immunosequencing reads came from different individuals (Table 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 5 provides details of the downloaded data and information about validated variations.

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 <i>mus musculus</i>	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)

Table 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.

Usage of D genes in the Healthy PBMC datasets. 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 7). 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 1) datasets (Figure 8). The novel variation N_Var(IGHD3-10*01)-3 appears in 7 datasets and indicates the heterozygous state of the gene in those datasets. It appears with the known variation IGHG3-10*01 accounting for 30–70% usage of the gene IGHG3-10 in those datasets.

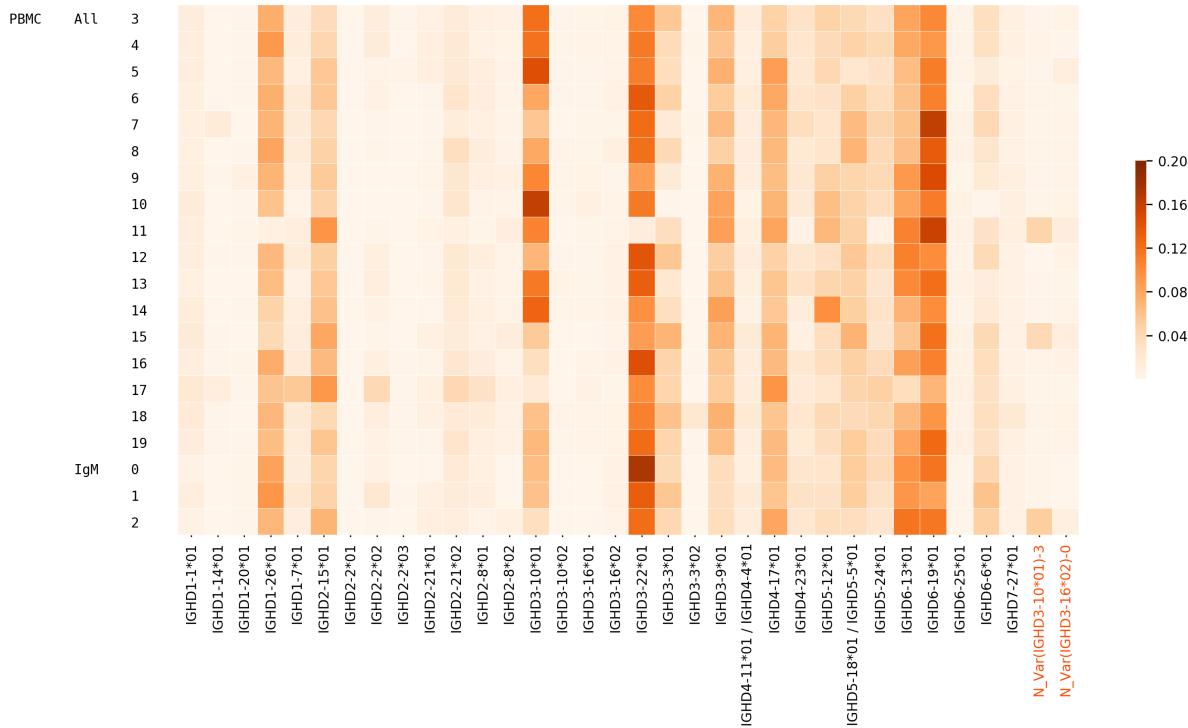


Figure 7. Usage of various known and novel genes in various Healthy datasets. Each row corresponds to a different dataset described by the three leftmost columns. The first column denotes the tissue, i.e., PBMC, the second column denotes the isotype (IgM or all/unsorted), and the numbers in the third 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.

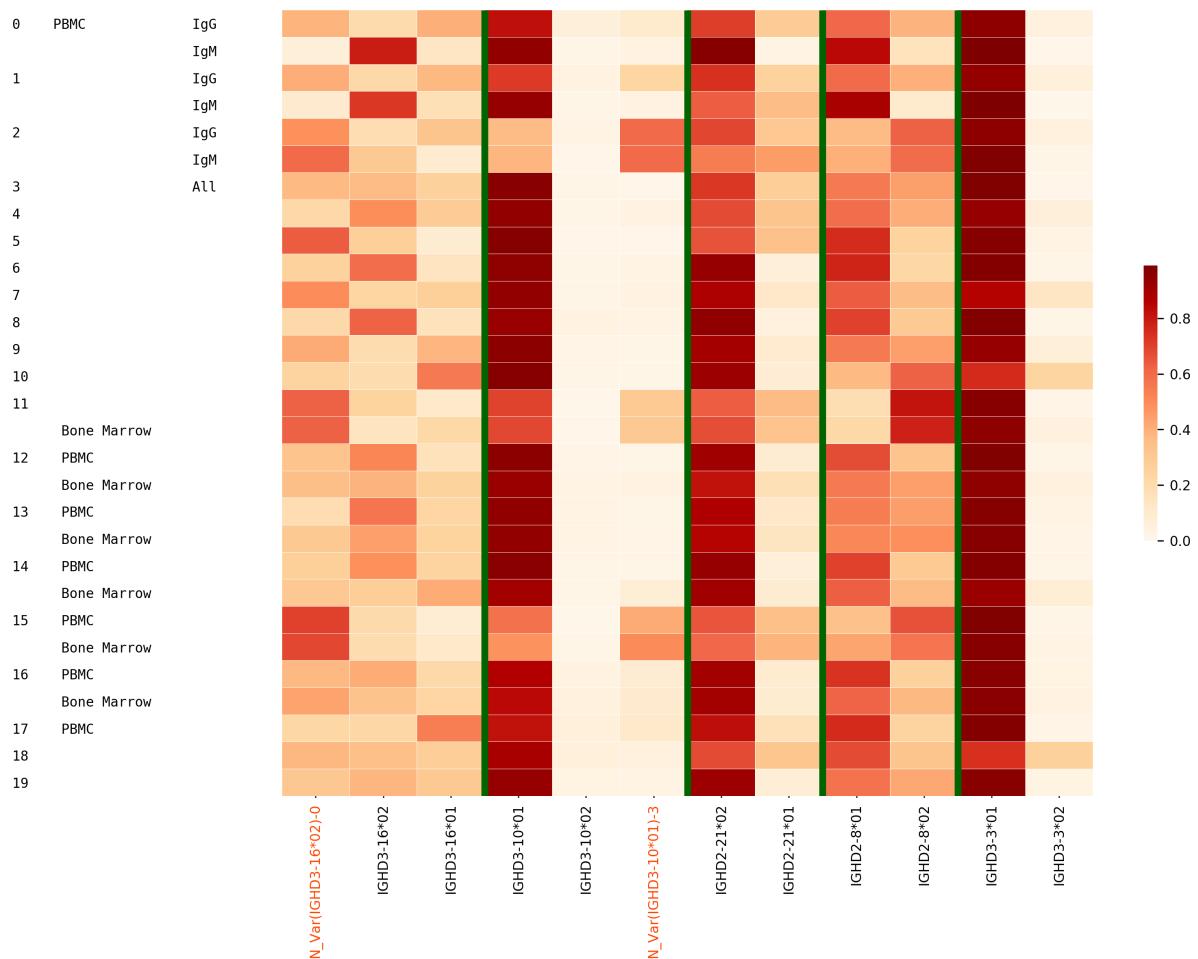


Figure 8. Usage of variants of D genes in Healthy PBMC BM datasets. Green lines separate the plot such that each subplot corresponds to one gene and its variants. Each cell in a subplot represents the proportion of the usage of a variant with respect to the total usage of all variants. Thus, in every 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. Since the total usage of the genes IGHD3-16 and IGHD2-8 is much less than usage of the genes IGHD3-10, IGHD2-21, and IGHD3-3 (as seen in Figure 7), it is difficult to comment on whether these genes are present in homozygous or heterozygous states.

Overused D genes in datasets specific to a health condition, tissue, and/or cell type. 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 9. Details on the genes overused in the Flu Vaccination datasets are shown in Table 6, 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 10), and 6 genes were overused in at least 2 datasets from the Intestinal datasets (Figure 11). 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.

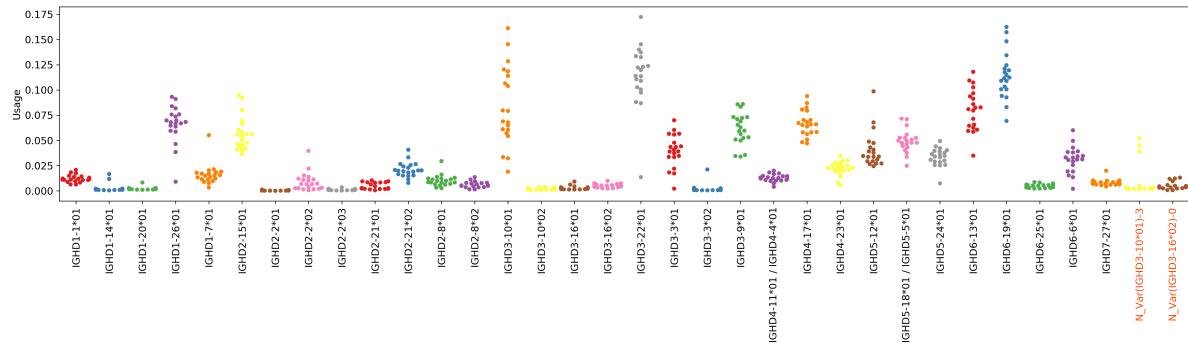


Figure 9. D gene usage in all Healthy datasets. Each point above a gene represents a Healthy Human PBMC dataset. To distinguish usages of different genes, adjacent genes are represented by different colors.

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

Table 6. 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.

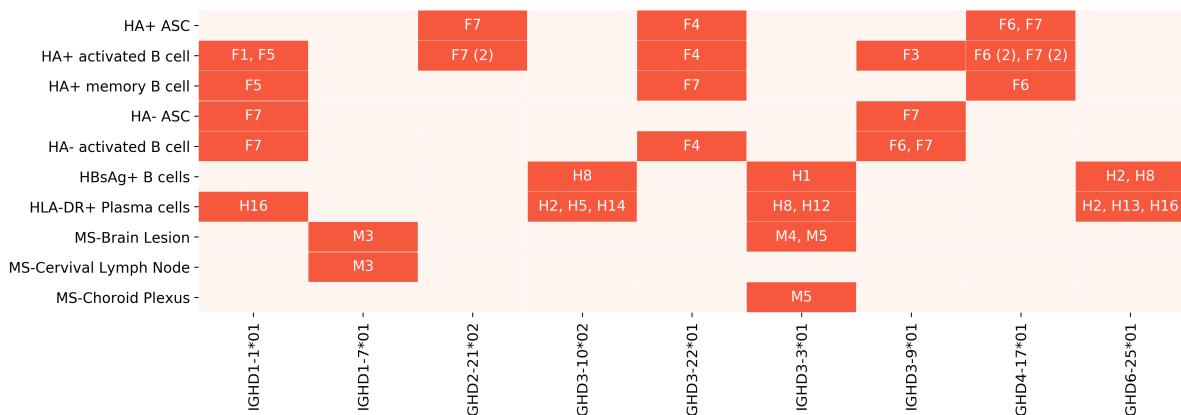


Figure 10. Summary of overused genes in Stimulated datasets. The datasets in which each gene is overused are highlighted and annotated with the corresponding individuals. Subjects were prefixed with a letter corresponding to the project – “F” for Flu Vaccination, “M” for Multiple Sclerosis, and “H” for 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.

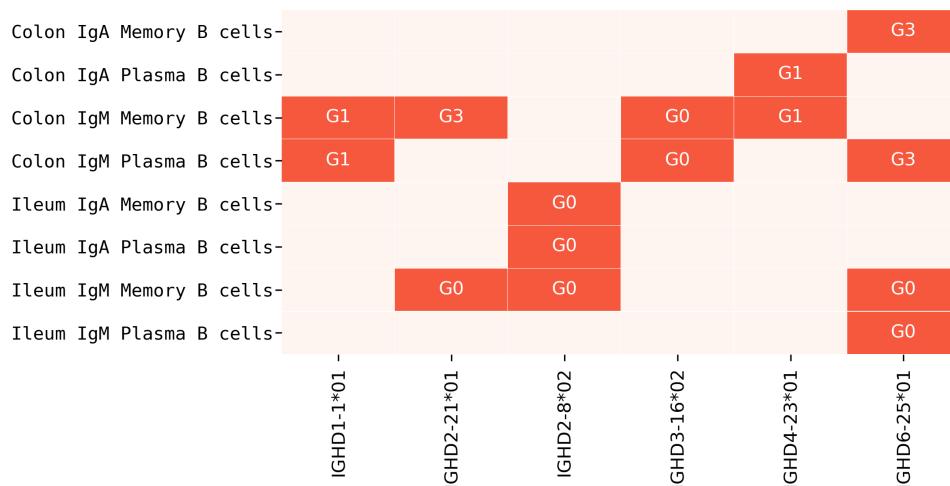


Figure 11. Summary of overused genes in Intestinal datasets. The datasets in which each gene is overused are highlighted and annotated with the corresponding individuals. The subjects in the Intestinal Repertoire project were prefixed with “G.”

Usage of D genes in the Mouse datasets. 57.4% of CDR3s on average were traceable in each dataset. Figure 12 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 13. 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.

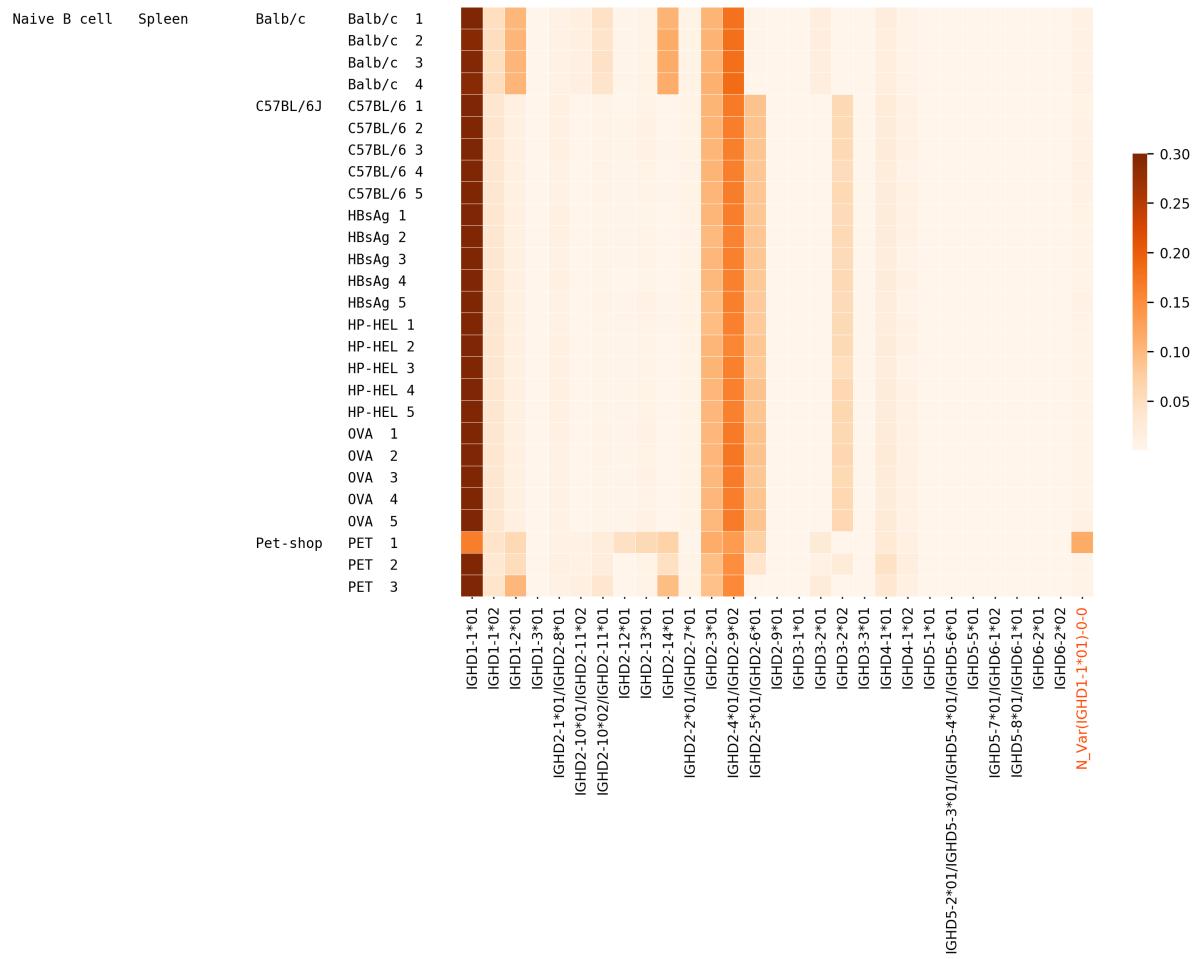


Figure 12. Usage of various known and novel genes/variations in MICE datasets. Columns on the left represent cell type, tissue, strain, and individual respectively. OVA, HP-HEL, and HBsAg in the right most column represent the C57BL/6J mice immunized with OVA, HP-HEL, and HBsAg, respectively. For example, OVA 3 represents the C57BL/6J mouse number 3 that was immunized with OVA.

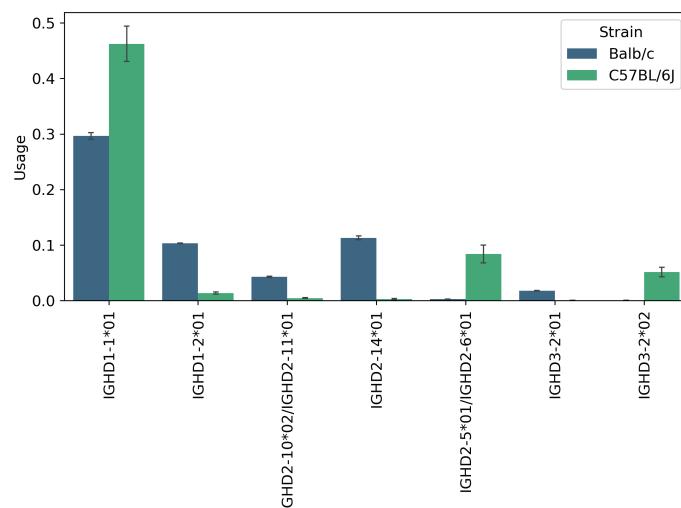


Figure 13. Genes with differential usage in Balb/c and C57BL/6J strains. Except IGHD1-1*01, all genes were inferred only in one strain.

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 S 12). For rats, genes belonging to the IGHD2 and IGHD3 families were used much less than in other gene families (Figure S 13). D genes with the highest usage among datasets of a species are shown in Supplemental Note: Highly Used D Genes in Non-human Datasets.

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).

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.

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.

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.

Availability of data and materials. MINING-D and all datasets analyzed in this paper are publicly available.

Competing interests. The authors declare no competing financial interests.

Funding. V.B. was supported by the Qualcomm Institute at UCSD. Y.S. was supported by the Data Science Fellowships at UCSD and the AAI Intersect Fellowship 2019. P.A.P. was supported by the NIH 2-P41-GM103484PP grant.

Author contributions. V.B. implemented the MINING-D algorithm. V.B. and Y.S. designed the computational experiments and performed benchmarking. V.B., Y.S., M.F., R.R., and P.A.P. conceived the study and wrote the manuscript.

References

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: PMC4136508.
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: PMC5595925.
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.co.2013.07.007. PubMed PMID: 23992996; PubMed Central PMCID: PMC4811628.
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: PMC4734744.
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: PMC4754645.
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: PMC2686541.
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: PMC4528413.
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.co.2016.02.003. PubMed PMID: 26922135; PubMed Central PMCID: PMC4884505.
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: PMC6391684.
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: PMC4281569.
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: PMC4345584.
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: PMC5187446.
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: PMC5095119.
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: PMC6381938.
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: PMC6516046.
29. 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: PMC6431624.
30. Mitzenmacher M. A survey of results for deletion channels and related synchronization channels. *Probability Surveys.* 2009;6:1-33.
31. 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.
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: PMC5054979.
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.jimmuni.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, Silleveld 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, Safanova 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.

Supplemental Notes

Greedy Algorithm

Defining Relative Positions

Removing Unidirectional Extensions

Immunosequencing Datasets

Novel Variations

MINING-D Parameters

D Gene Usage

Overused D Genes

Highly Used D Genes in

Supplemental Note: 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 S.1 Pseudocode of the greedy algorithm

Supplemental Note: 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 gene segments, 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 segment, 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 S 2 and Figure S 3.

Figure S 2. 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 CGAAATACCA is 0.32, whereas the relative position of its potential extension in red is 0.04.

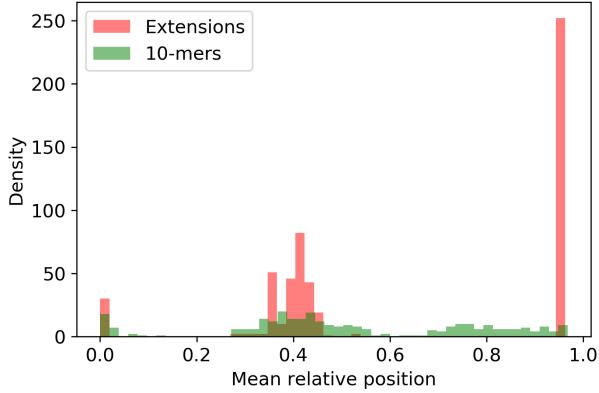


Figure S 3. The mean relative positions of the abundant seed 10-mers (in green) and their extensions (in red) in the MOUSE dataset. The relative positions of the extensions form three clusters, each corresponding to one of the V, D, and J gene segments.

Supplemental Note: 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 segment. For example, the k -mer shown in Figure S 4 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 segment 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.

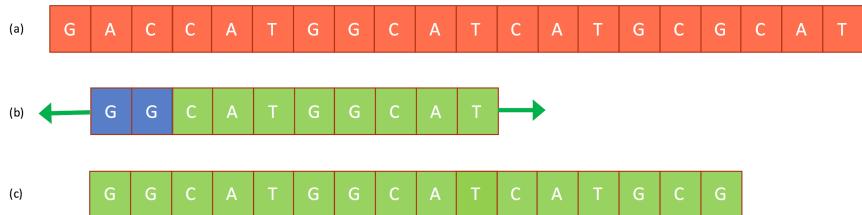


Figure S 4. A highly abundant 10-mer (b) that is formed by random insertions (two nucleotides in the beginning) and 8 nucleotides from a highly abundant D gene (a). Since this 10-mer was not a substring of the D gene, its extension (c) is also not a substring of the D gene.

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 S 1.

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

Table S 1. Possible values of N_L and N_R with the constraint when $\alpha = 0.5$.

Supplemental Note: Immunosequencing Datasets

Summaries of all the human and non-human immunosequencing datasets analyzed in this study are shown in Table S 2 and Table S 3, respectively.

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 S 2. Summary of human immunosequencing datasets analyzed in the study. ASC refers to antibody secreting cells.

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
--------	--------------------------	------------------------	---	--------------	----------	----	-------------	---

Table S 3. Summary of non-human immunosequencing datasets analyzed in the study.

Supplemental Note: Novel Variations

All the variations found using MINING-D for humans, camels, rhesus macaques, mice, rats, and rabbits are shown in Table S 4. The polymorphisms in the genes validated using genomic data are highlighted in red.

Human	
IGHD3-10*01 Original GTATTACTATGGTTGGGGAGTTATTATAAC N_Var-3 GTATTACTATGGTTCAAGGGAGTTATTATAAC N_Var-2 -----ATGGTTGGGGAGTTATTATAAC N_Var-1 -----TGGTTGGGGAAATT----- N_Var-0 -----TGaTTCGGGGAGTT-----	IGHD3-22*01 Original GTATTACTATGATACTAGTGGTTATTACTAC N_Var-1 ---ATTACTATGATACTAGTGG----- N_Var-0 -----TATGATACTAGTGGT-----
IGHD2-2*01 Original AGGATATTGTAGTAGTACCAAGCTGCTATGCC N_Var-0 AGGATATTGTATAGTACCAAGCTGCTAT---	IGHD3-16*02 Original GTATTATGATTACGTTGGGGAGTTATCGTTATACC N_Var-0 ---TTATGATTACATTGGGGAGTTATCGTTAT---
Camel	
IGHD3*01 (Alpaca) Original GTATTACTACTGCTCAGGCTATGGGTGTTATGAC N_Var-1 ---GACTGCTATTCAGGCTCTTGGTGTATG-- N_Var-0 ---TGACTACTGTTCAAGCTCTTGGTGT-----	IGHD2*01 (Alpaca) Original ACATACTATAGTGGTAGTTACTACACC N_Var-1 --ATATGTTAGTGGTGGTTACTGCTAC--- 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 AGACTACGGGTTGGGTAC N_Var-0 ----TATGGGTT-GGGTAC	
Rhesus Macaque	
IGHD1S39*01 Original GGTATAGTGGGAACTACAAC N_Var-0 -----AGTGGGA G CTAC---	IGHD3S18*01 Original GTACTGGGTGATTATTATGAC N_Var-0 --ACTGG A GTGATTATTA---
IGHD5S3*01 Original GTGGATACAGTGGGTACAGTTAC N_Var-0 ---GATACAG C GGGTACAGT---	IGHD2S11*01 Original AGAATATTGTAGTAGTACTTACTGCTCCTCC N_Var-0 -----ATTGTAGTGGTACTTACTGCT----
IGHD2S17*01 Original AGAATACTGTTACTGGTAGTGTTGCTATGCC N_Var-0 ----TACTGTTACTGGTAGTGTTGCTAC---	IGHD3S23*01 Original GTATTACTATGATACTGGTTATTACACCCACACCGT N_Var-0 ---TTACTATG C TAGTGGTTATTAC-----
Mouse	
IGHD1-1*01 Original TTTTATTACTACGGTAGTAGCTAC- N_Var-3 -----ACgACGGTAGTAGC--- N_Var-2 -TTTATTACTACGGTAGTAGCAGgggg N_Var-1 ---ATTACTgCGGTAGTAGCTAC- N_Var-0 TTTTATTACTACGG A TGGTAGCTACg	IGHD2-4*01 Original TCTACTATGATTACGAC--- N_Var-0 -----ggGATTACGACagg
Rat	

IGHD1-3*01 Original TTTTTAACTATGGTAGCTAC N_Var-0 -TTTTAACTA C GGTAGCTAC	IGHD1-9*01 Original TACATACTATGGGTATAACTAC- N_Var-1 --CATACTA C GGGTATA C CTACg N_Var-0 --CATACTAcGGGTATAACTAC-
IGHD1-12*02 Original TTTATTACTATGATGGTAGTTTACTAC- N_Var-0 -TTTATTACTATGATGGTA C TTTACTACg	
Rabbit	
IGHD6-1*01 Original -----GTTACTATAGTTATGGTTATGCTTATGCTACC N_Var-7 -----TTAtgATggTTATGGTTATGgTa----- N_Var-6 tta-----tagtgGTTAtggTgGTTATGcTTATG----- N_Var-5 -----TggTTATgTgATGGTTATGCT----- N_Var-4 TTACTATACTATGGTTATGC TGGTTATGCT AC ----- N_Var-3 TTA ----- TGCG GTTAT GC TGGTTATGCT AC ----- N_Var-2 -----tgGTTAtggTgGTTATGCT----- N_Var-1 -----ATActTATGGTTATGgTggT----- N_Var-0 -----ATAGTTATGGTTATGgTg-----	IGHD1-1*01 Original GCATATACTAGTAGTAGTGTTATTATATAC N_Var-2 GCATAT G CTAGTAGTAGTGTTATTAT----- N_Var-1 -----TgTAGTGTTATTAT----- N_Var-0 -----GTAGTGTTggTTAT-----
IGHD2-1*01 Original TAGCTACGATGACTATGGTATTAC- N_Var-0 -----TGatTATGGTggTTAtg	IGHD8-1*01 Original GTTATGCTGGTAGTAGTTATTATACC N_Var-0 -TTATGCTGGTgaTgGTTATg-----

Table S 4. All Inferred novel variations. Variations that were validated using genomic data are shown by highlighting the polymorphisms in red.

Supplemental Note: 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 S 5. 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.

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)

Table S 5. 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.

Supplemental Note: 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 S 5). 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.

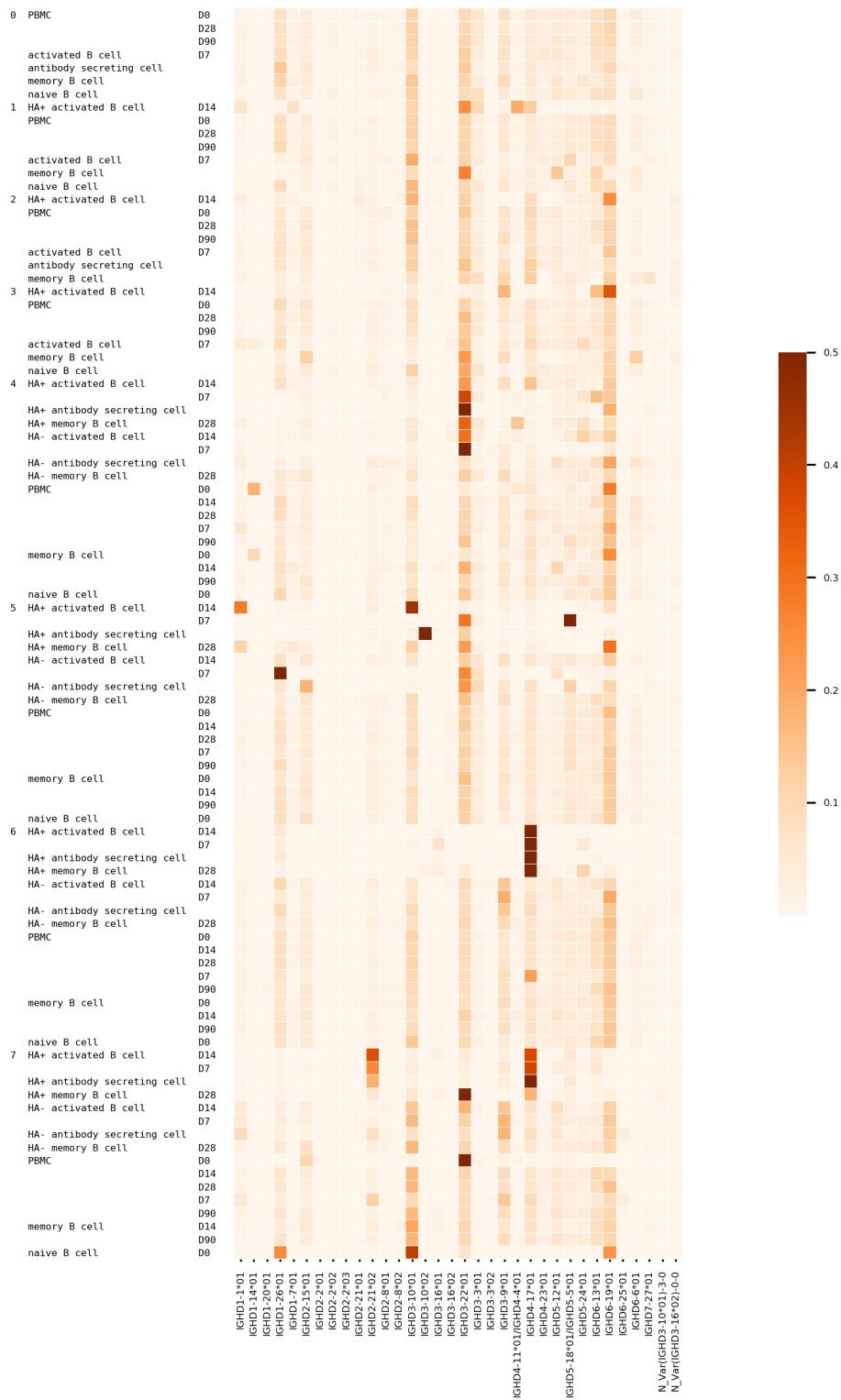


Figure S 5. Usage of IMGT and novel variations of IGHD genes in various datasets corresponding to flu vaccination. The columns on the left represent the individual, the cell type, and the time point (day after vaccination).

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 S 6. 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.

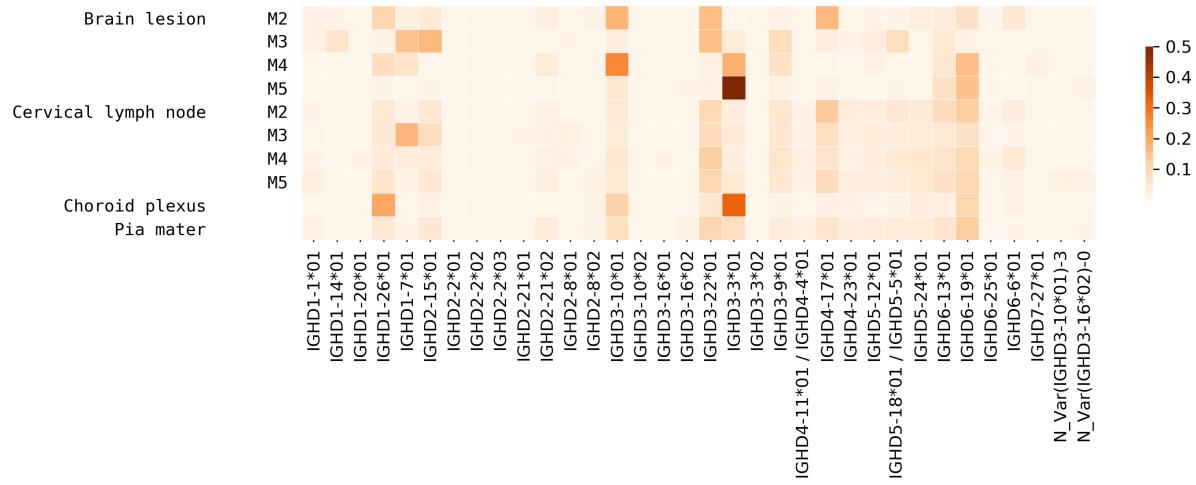


Figure S 6. Usage of various known and novel genes in various datasets corresponding to different tissues in Multiple Sclerosis patients. The columns on the left represent the tissue and the individual, respectively.

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 S 7). 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 7). 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.

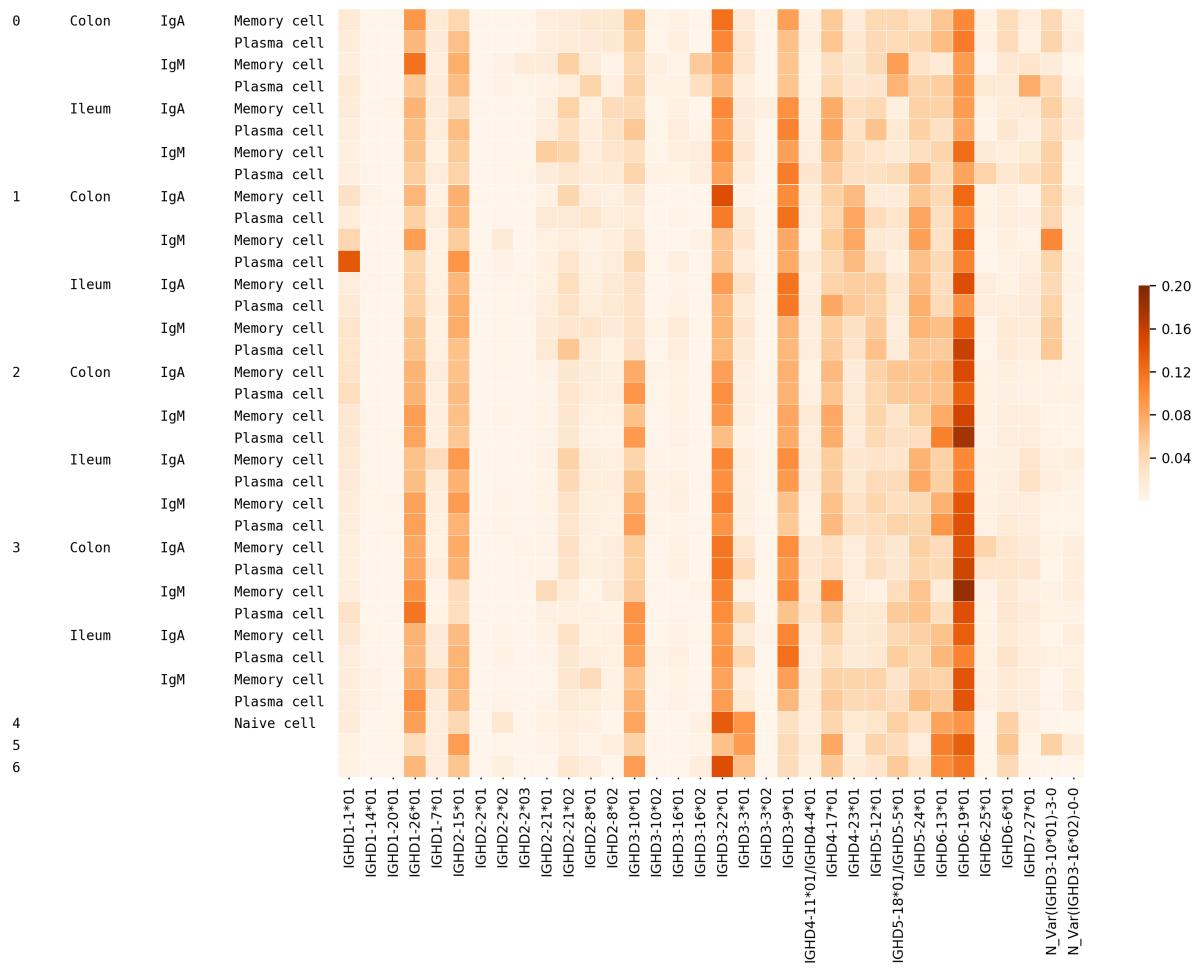


Figure S 7. Usage of various known and novel genes in various datasets corresponding to human intestinal antibodies. The columns on the left represent the individual, tissue, isotype, and cell type, respectively.

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 S 8. 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. Comparing the usage in HbsAg+ B cells and HLA-DR+ plasma cells, the gene IGHD3-22*01 seems to be unpresented, and the genes IGHD5-5*01/IGHD5-18*01 and IGHD7-27*01 are overused in HLA-DR+ plasma 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.

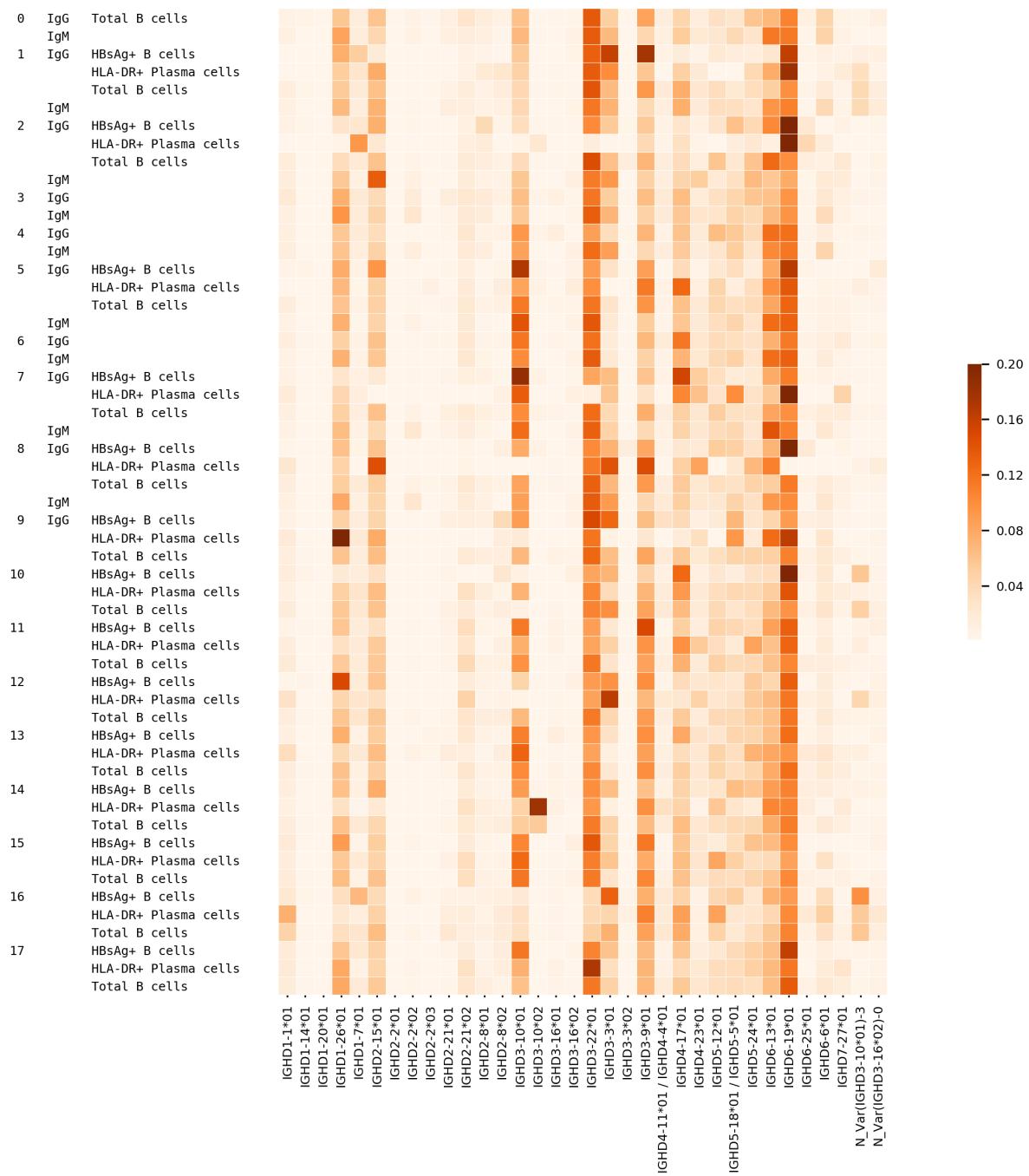


Figure S 8. Usage of various known and novel genes in different datasets corresponding to different cell types and isotypes corresponding to human subjects with hepatitis B vaccination.

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 S 9).

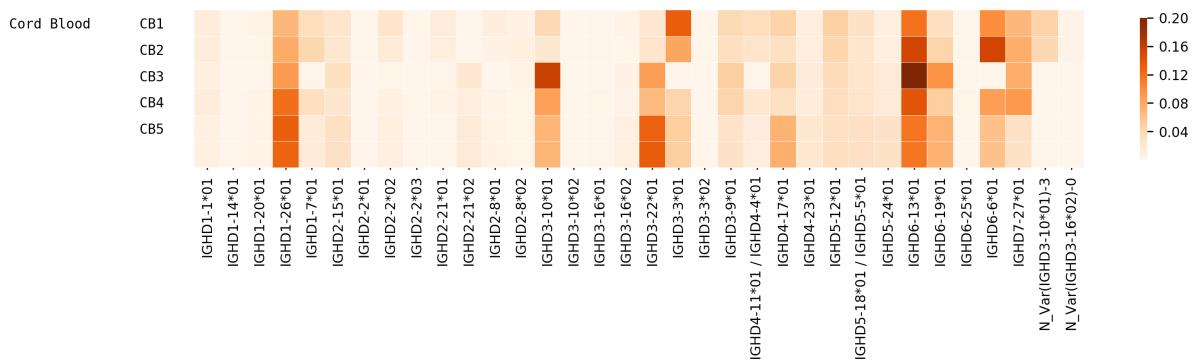


Figure S 9. Usage of various known and novel genes in different PBMC and cord blood datasets.

Usage of D genes in mice datasets.

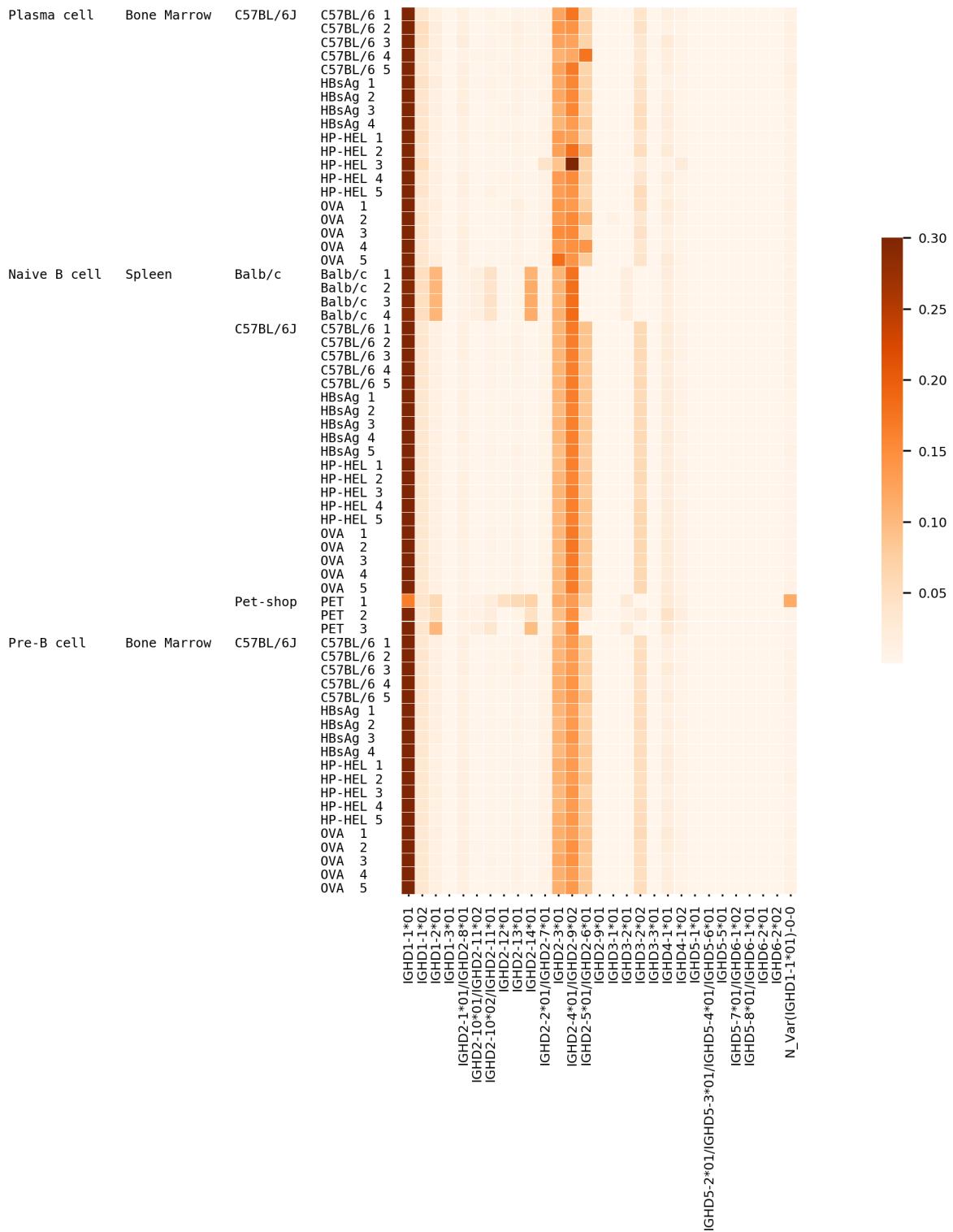


Figure S 10. Usage of various known and novel genes/variations in different datasets corresponding to different strain, cell type, and tissue from mice. Columns on the left represent cell type, tissue, strain, and individual, respectively. OVA, HP-HEL, and HBsAg in the right most column represent the 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.

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 S 11.

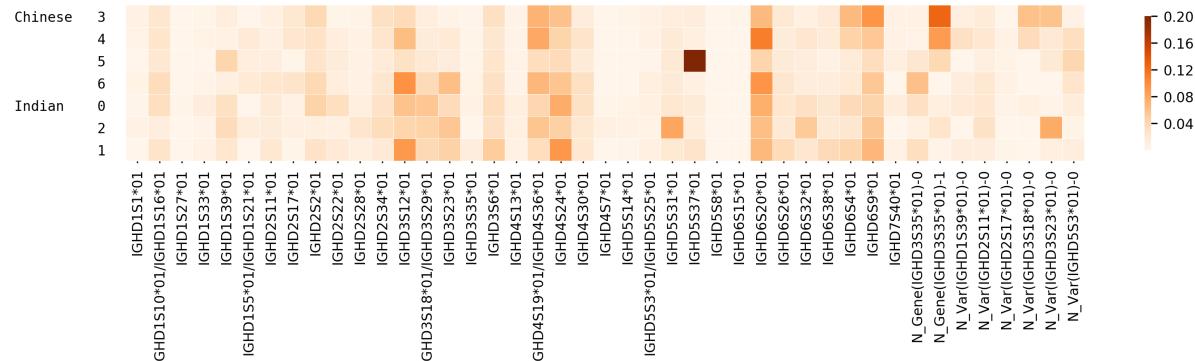


Figure S 11. Usage of known and novel genes in the Rhesus Macaque datasets. The novel genes and variations are shown on the right.

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 S 12. 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.

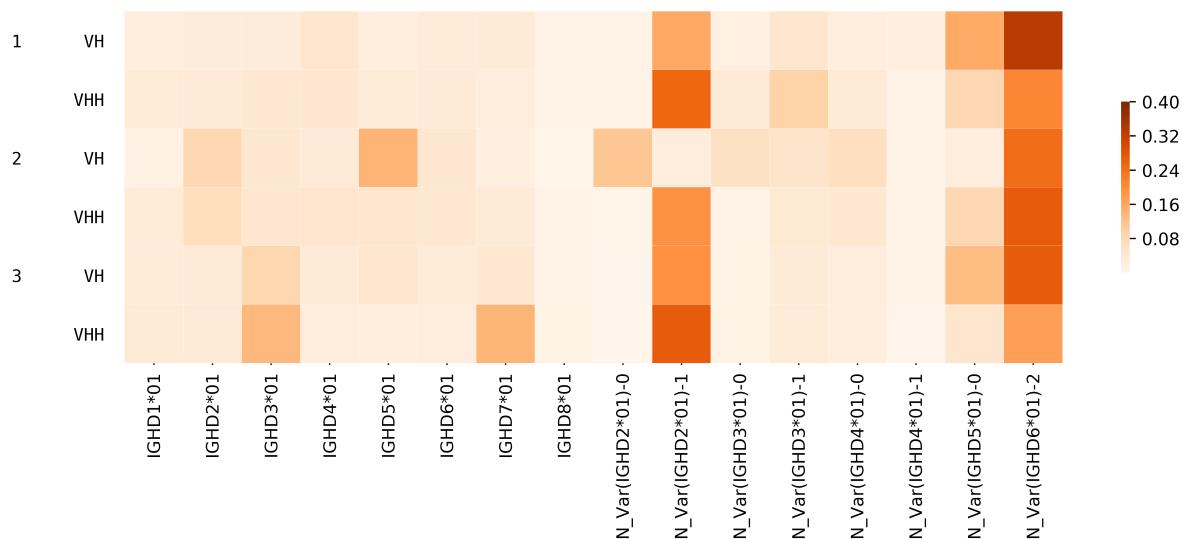


Figure S 12. Usage of known and novel genes in the Camel datasets.

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 S 13. 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.

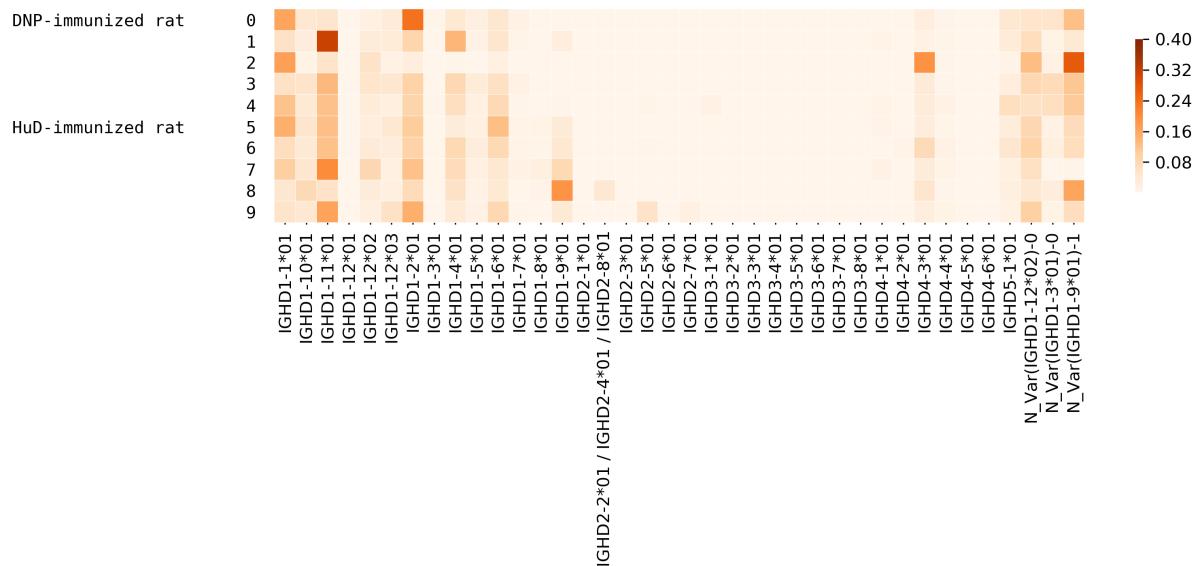


Figure S 13. Usage of D genes in the Rat datasets.

Supplemental Note: Overused D genes

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
		Brain Lesion	8.2x
	M5	Choroid Plexus	4.7x

Table S 6. Overused genes in the Multiple Sclerosis datasets.

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

Table S 7. Overused genes in the Intestinal Repertoire 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
				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 S 8. Overused genes in the Hepatitis B vaccination datasets

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

Table S 9. Overused genes in the Cord Blood datasets.

Supplemental Note: 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 S 10 and Figure S 14).

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

Table S 10. 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.

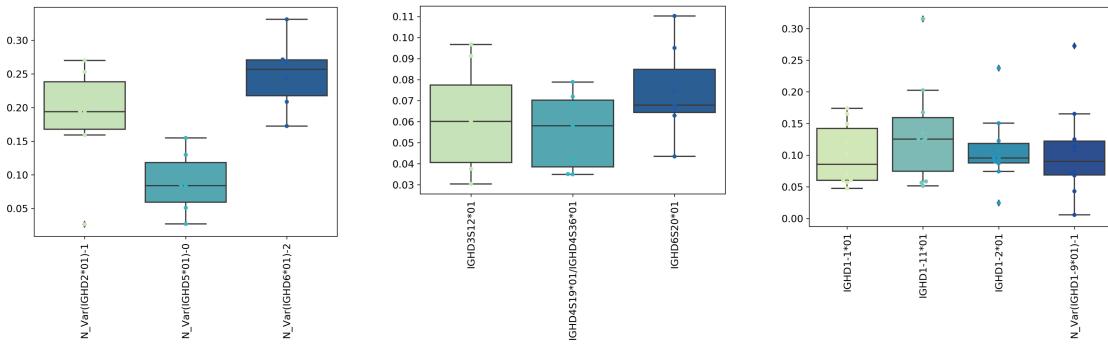


Figure S 14. Usage proportion of highly used genes in the Camel (left), Macaque (middle), and Rat (right) datasets.

Supplemental Note: IgScout vs. MINING-D

We compared the results of IgScout and MINING-D on all datasets from the project PRJEB18926. The results are shown in Figure S 15. 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. IgScout failed to reconstruct 7 genes in all the datasets whereas MINING-D failed to reconstruct only 3 genes in all the datasets.

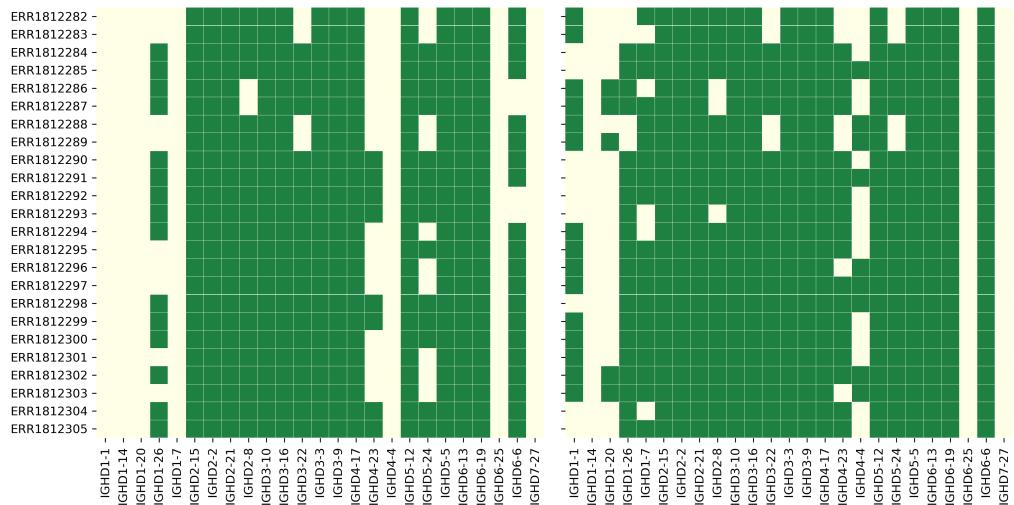


Figure S 15. Results of IgScout (left) and MINING-D (right) on datasets from the project PRJEB18926. All genes that were found in a dataset are shown in dark green, whereas the missing genes in datasets are denoted by light green.