



# De novo Inference of Diversity Genes and Analysis of Non-canonical V(DD)J Recombination in Immunoglobulins

Yana Safonova<sup>1\*</sup> and Pavel A. Pevzner<sup>2</sup>

<sup>1</sup> Center for Information Theory and Applications, University of California, San Diego, San Diego, CA, United States,

<sup>2</sup> Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA, United States

## OPEN ACCESS

### Edited by:

Deborah K. Dunn-Walters,  
University of Surrey, United Kingdom

### Reviewed by:

Richard L. Frock,  
Stanford University, United States  
Mats Ohlin,  
Lund University, Sweden

### \*Correspondence:

Yana Safonova  
isafonova@eng.ucsd.edu

### Specialty section:

This article was submitted to  
B Cell Biology,  
a section of the journal  
Frontiers in Immunology

**Received:** 17 January 2019

**Accepted:** 16 April 2019

**Published:** 03 May 2019

### Citation:

Safonova Y and Pevzner PA (2019) De  
novo Inference of Diversity Genes and  
Analysis of Non-canonical V(DD)J  
Recombination in Immunoglobulins.  
Front. Immunol. 10:987.  
doi: 10.3389/fimmu.2019.00987

The V(D)J recombination forms the immunoglobulin genes by joining the variable (V), diversity (D), and joining (J) germline genes. Since variations in germline genes have been linked to various diseases, personalized immunogenomics aims at finding alleles of germline genes across various patients. Although recent studies described algorithms for *de novo* inference of V and J genes from immunosequencing data, they stopped short of solving a more difficult problem of reconstructing D genes that form the highly divergent CDR3 regions and provide the most important contribution to the antigen binding. We present the IgScout algorithm for *de novo* D gene reconstruction and apply it to reveal new alleles of human D genes and previously unknown D genes in camel, an important model organism in immunology. We further analyze non-canonical V(DD)J recombination that results in unusually long CDR3s with tandem fused IGHD genes and thus expands the diversity of the antibody repertoires. We demonstrate that tandem CDR3s represent a consistent and functional feature of all analyzed immunosequencing datasets, reveal ultra-long CDR3s, and shed light on the mechanism responsible for their formation.

**Keywords:** repertoire sequencing, VDJ recombination, germline gene inference, antibody repertoire, repertoire diversity

## INTRODUCTION

Antibodies provide specific binding to an enormous range of antigens and represent a key component of the adaptive immune system. The *antibody repertoire* is generated by *somatic recombination* of the V (*variable*), D (*diversity*), and J (*joining*) germline gene segments. Immunosequencing has emerged as a method of choice for generating millions of reads that sample antibody repertoires and provide insights into monitoring immune response to disease and vaccination (1).

Information about all germline genes in an individual is a pre-requisite for analyzing immunogenomics data. However, nearly all immunogenomics studies rely on the population-level germline genes rather than germline genes in a specific individual that the immunosequencing data originated from. 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 (2, 3). Moreover, it is non-trivial to figure out 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 (3). These errors hide the identity of the individual germline genes, make it difficult to analyze *somatic hypermutations* (SHM) and complicate studies of antibody evolution (4–6).

*Personalized immunogenomics* (i.e., identifying individual germline genes) is important since variations in germline genes have been linked to various diseases (7), differential response to infection, vaccination, and drugs (8, 9), aging (10), and disease susceptibility (7, 11, 12). However, since the International ImmunoGeneTics (IMGT) database is incomplete even in the case of well-studied human germline genes (13), there exist still unknown human allelic variants that are difficult to differentiate from SHMs. In the case of immunologically important but less studied model organisms, such as camels or sharks, the germline genes remain largely unknown. Unfortunately, since assembling the highly repetitive immunoglobulin locus from whole genome sequencing data faces challenges (14), the efforts like the 1,000 Genomes Project have resulted only in limited progress toward inferring the population-wide census of germline genes (14–16).

In addition to personalized immunogenomics, the incompleteness of the IMGT database negatively affects analysis of monoclonal antibodies. Existing tools for antibody sequencing from tandem mass spectra (17, 18) rely on a comprehensive database of V, D, and J genes to assemble tandem mass spectra into an intact antibody. Lack of such databases for many species limits applications of Valens (Digital Proteomics), SuperNova (Protein Metrics), and other software tools for antibody sequencing.

Although the personalized immunogenomics approach was first proposed by Boyd et al. (19), the manual analysis in this study did not result in a software tool for inferring germline genes. Gadala-Maria et al. (20) developed the TIgGER algorithm for inferring germline genes and used it to discover 11 novel allelic V segments. However, 20 stopped short of *de novo* reconstruction of the germline genes and acknowledged that it is important to develop algorithms for finding diverged alleles that TIgGER is not able to find. In the case of V and J genes, this challenge was addressed by Corcoran et al. (21), Zhang et al. (22), and Ralph and Matsen (3). However, as Ralph and Matsen (3) commented, the more challenging task of *de novo* reconstruction of D genes remains elusive. This is unfortunate since D genes contribute to the *complementarity determining region 3* (CDR3) that covers the junctions between V, D, and J genes and represents the highly divergent part of antibodies. We describe the IgScout algorithm for *de novo* inference of D genes and apply it to diverse immunosequencing datasets with the goal to reconstruct dominant variants of highly abundant D genes and discover novel highly abundant variations.

Although many studies analyzed patterns of V-D-J pairing (23, 24), there is still a shortage of studies of unusual recombination events such as *V(DD)J recombination* incorporating two D genes into a single unusually long CDR3 with tandem fused IGHD genes (or *tandem CDR3*). Meek et al. (25) were the first to reveal a few tandem CDR3s, thus confirming the *V(DD)J recombination* conjecture put forward by Kurosawa and Tonegawa (26). However, since tandem CDR3s are rare, they remained elusive for the next two decades and (27, 28) even argued that tandem

CDR3s found in Meek et al. (25) represent artifacts. However, Briney et al. (29) and Larimore et al. (30) demonstrated that tandem CDR3s do exist (at frequency 1 per 800 B-cells) by analyzing high-throughput immunosequencing datasets.

As emphasized in Briney et al. (29), detecting *V(DD)J recombination* has to be done with caution since it is often confused with standard *V(D)J recombination*. Although they came up with a heuristic for detecting tandem CDR3s, there is still no software for detecting tandem CDR3s and it remains unclear how many tandem CDR3s found in Briney et al. (29) represent false positives. We thus extended the functionality of the IgScout algorithm to finding tandem CDR3s and revealed that *V(DD)J recombination* is a functional (rather than aberrant) feature with frequency varying from 1 per 200 to 1 per 2,500 B-cells across various datasets. Finally, we revealed *ultra-long tandem CDR3s* and shed light on the mechanism responsible for their formation.

## RESULTS

### Immunosequencing Datasets

We analyzed the following datasets described in the **Supplemental Note** “Immunosequencing datasets”:

- **HEALTHY**: 14 datasets from 14 healthy human donors,
- **ALLERGY**: 24 datasets from six allergy patients (31),
- **HIV**: 13 datasets from two HIV-infected patients (32),
- **NAÏVE**: 7 datasets from naïve B cells of healthy human donors,
- **PROJECT10**: 600 datasets from various humans resulting from 10 NCBI projects
- **CAMEL**: 6 datasets from three healthy camels (33).

### Constructing CDR3 Datasets

We illustrate the work of IgScout using one of the HEALTHY datasets (Set 1) containing heavy chain repertoires extracted from *peripheral blood mononuclear cells* (PBMC). The IgReC tool (34) extracted 228,619 distinct CDR3s from this dataset. To minimize impact of sequencing and amplification errors, we clustered similar CDR3s (differing by at most three mismatches) and constructed consensus for each cluster resulting in 98,576 *consensus CDR3* of average length 46 nucleotides.

Each CDR3s typically starts from a short suffix of a V gene and ends with a short prefix of a J gene. Since these suffixes and prefixes negatively affect reconstruction of D genes, IgScout trims them as described in the **Supplemental Note** “Preprocessing CDR3 datasets.” This procedure reduces the average length of CDR3 strings (46 nucleotides) to 30 nucleotides strings that represent substrings of CDR3s that are not encoded by IGHV or IGHJ genes. The result of the procedure is the set of strings *CDR3\**. We refer to the number of strings in *CDR3\** as  $|CDR3^*|$ .

### Overview of Human D Genes

The human immunoglobulin (IGH) locus contains 27 D genes that vary in length from 11 to 37 nucleotides. Since two pairs of human D genes are identical, there exist only 25 distinct D genes. Since the IMGT database refers to D genes using rather

long names and since these names do not reveal the ordering of D genes in the IGH loci (that is important for analyzing tandem CDR3s), it is difficult to visualize the IgScout results across all D genes and across multiple immunosequencing datasets. We thus renamed distinct human D genes from D1 to D27 in the increasing order of their positions in the IGH locus. The IMGT database also contains seven alleles of D genes denoted D2\*2, D2\*3, D3\*2, D8\*2, D10\*2, D16\*2, and D21\*2. See **Table 1** and **Supplemental Note** “Information about human D genes” for details.

## Frequent *k*-mers in D Genes

The problem of inferring germline genes can be formulated as the Trace Reconstruction Problem (35) in information theory described in the Methods section. IgScout is a heuristic for solving this problem that is inspired by the RepeatScout algorithm for *de novo* repeat finding (36) and that is based on analyzing frequent *k*-mers (contiguous strings of length *k*) in CDR3s. We illustrate the work of IgScout using *k*-mers of size 15 (all human D genes are longer than 15 nucleotides except for 11 nucleotide long gene D27).

The human D genes contain 305 15-mers. We classify a *k*-mer as *known* if it occurs in a human D gene (from IGHD1-1 to IGHD7-27), *mutated* if it differs from a known *k*-mer by a single substitution, and *trimmed* if it contains a known (*k*-2)-mer. All other *k*-mers are called *foreign*. Twenty-seven percent of strings in the CDR3\* dataset contain a known 15-mer and 35% contain either a known, or a mutated, or a trimmed 15-mer.

We classify a *k*-mer as *common* if its abundance exceeds  $\text{fraction}^* |\text{CDR3}^*|$  (the default value  $\text{fraction}=0.001$ ). **Figure 1** and the **Supplemental Note** “Common *k*-mers” present distributions of frequencies of all common 15-mers in various datasets. Although the vast majority of common *k*-mers are known, mutated, or trimmed, some of them are foreign. These

foreign common *k*-mers have to be treated with caution since they may trigger false positive inferences of D genes.

## From Frequent *k*-mers to D Gene Reconstruction

IgScout selects a most abundant *k*-mer in the CDR3\* dataset, aligns all CDR3 that contain this *k*-mer (using this *k*-mer as the alignment seed), and constructs the *motif logo* of the resulting alignment (**Figure 1**). It further trims all positions of the motif logo with the *information content* below *IC* (the default value  $IC = 0.5$ ) and computes the consensus string. Afterwards, it extends the consensus strings to the right and to the left (the PrefixExtension and SuffixExtension steps in the **Supplemental Note** “IgScout pseudocode”) to construct a putative D gene as described in the Methods section. Finally, the algorithm removes the sequences that contain *k*-mers from the identified putative D gene from the set CDR3\*, finds a most abundant *k*-mer in the resulting dataset, and iterates. IgScout stops when a most abundant *k*-mer is not a common *k*-mer (see **Supplemental Notes** “IgScout pseudocode,” “IgScout parameters,” and “Benchmarking IgScout on simulated immunosequencing datasets”). **Figure 2** demonstrates that IgScout reconstructs many known human D genes.

Similarly to the existing tools for reconstructing V and J genes (that typically trim a few nucleotides in the beginning/end of the reconstructed genes), IgScout also trims a few nucleotides in the beginning/end of the reconstructed D genes. Although lowering the *IC* threshold would reduce the number of trimmed nucleotides, we decided not to do it since lowering this parameter may result in erroneous reconstructions and since the trimmed nucleotides hardly affect the downstream applications of IgScout. See **Supplemental Note**: “How trimmed (rather

**TABLE 1** | Positions and lengths of human D genes.

| Name | IMGT name | Position (bp) | Length (nt) | Name | IMGT name | Position (bp) | Length (nt) |
|------|-----------|---------------|-------------|------|-----------|---------------|-------------|
| D1   | IGHD1-1   | 105,919,502   | 17          | D15  | IGHD2-15  | 105,897,957   | 31          |
| D2   | IGHD2-2   | 105,916,826   | 31          | D16  | IGHD3-16  | 105,895,634   | 37          |
| D3   | IGHD3-3   | 105,914,359   | 31          | D17  | IGHD4-17  | 105,894,508   | 16          |
| D4   | IGHD4-4   | 105,913,222   | 16          | D5   | IGHD5-18  | 105,893,542   | 20          |
| D5   | IGHD5-5   | 105,912,257   | 20          | D19  | IGHD6-19  | 105,891,699   | 21          |
| D6   | IGHD6-6   | 105,910,410   | 18          | D20  | IGHD1-20  | 105,891,191   | 17          |
| D7   | IGHD1-7   | 105,909,907   | 17          | D21  | IGHD2-21  | 105,888,551   | 28          |
| D8   | IGHD2-8   | 105,907,211   | 31          | D22  | IGHD3-22  | 105,886,031   | 31          |
| D9   | IGHD3-9   | 105,904,681   | 31          | D23  | IGHD4-23  | 105,884,870   | 19          |
| D10  | IGHD3-10  | 105,904,497   | 31          | D24  | IGHD5-24  | 105,883,903   | 20          |
| D4   | IGHD4-11  | 105,903,616   | 16          | D25  | IGHD6-25  | 105,881,539   | 18          |
| D12  | IGHD5-12  | 105,902,649   | 23          | D26  | IGHD1-26  | 105,881,034   | 20          |
| D13  | IGHD6-13  | 105,901,142   | 21          | D27  | IGHD7-27  | 105,865,551   | 11          |
| D14  | IGHD1-14  | 105,900,638   | 17          |      |           |               |             |

Since the IGH locus starts at the end of the 14th chromosome, positions are given with respect to its complementary sequence (assembly GRCh38.p12). Green and orange cells correspond to two duplicated and identical D genes IGH4-4\*01–IGH4-11\*01 (D4) and IGH5-5\*01–IGH5-18\*01 (D5).

than complete) D genes affect the downstream analysis of immunosequencing datasets.”

Indeed, the personalized immunogenomics applications [such as the discovery of “deficient” germline variants that lead to poor responses to vaccination (12)] are hardly affected by the fact that all existing tools for inferring the V, D, and J genes trim a few nucleotides from the ends. Reconstruction of monoclonal antibodies from tandem mass spectra and various proteogenomics applications are also hardly affected by this trimming. Moreover, in the case of human germline genes (and other genomes with well-characterized germline genes) the trimmed nucleotides can be tentatively reconstructed based on similarity with known germline genes (as has been done in previous studies of V and J genes). However, in some cases, assigning terminal nucleotides by homology might lead to the inference of erroneous alleles (38–40). Ideally, the gene inference problem should be followed by validation using genomic data that raises need in paired Rep-Seq and WGS datasets from the same individual. The antibody analysis and engineering in model organisms can also be done with partial D genes.

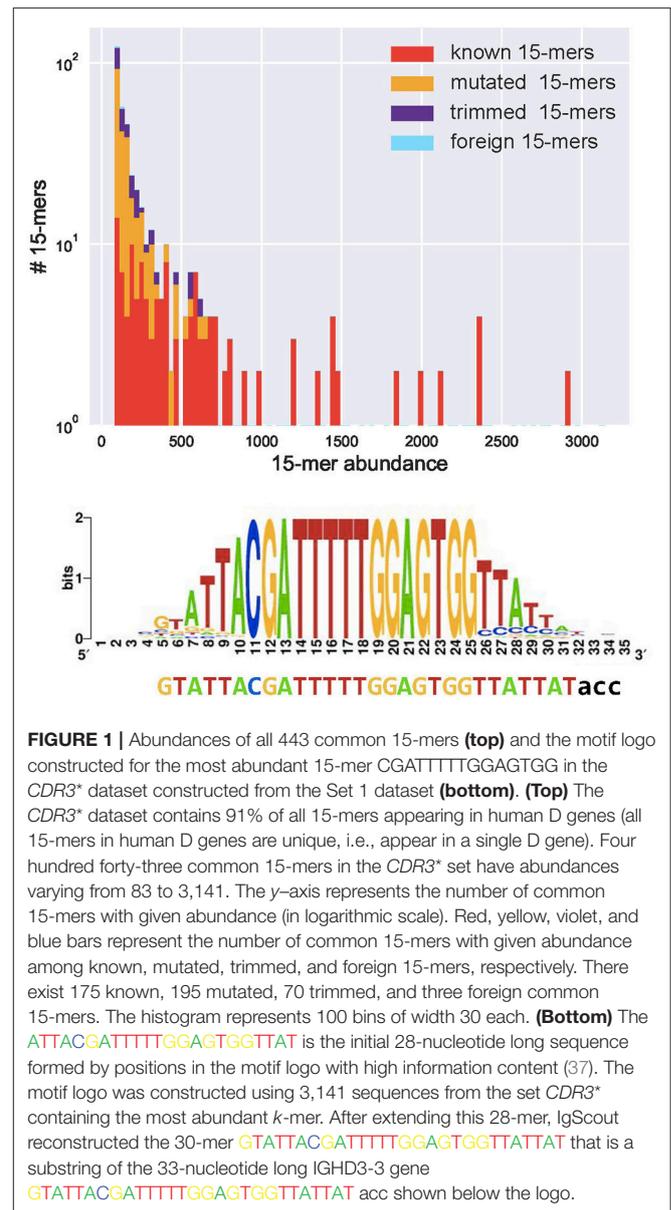
## Limitations and Advantages of IgScout

The IgScout pipeline consists of three steps: (i) preprocessing Rep-seq reads; (ii) inferring D genes; (iii) analyzing VDJ recombinations based on the inferred genes (Figure 3). The preprocessing step extracts CDR3s, constructs consensus CDR3s, and trims prefixes and suffixes of CDR3s to exclude suffixes of V genes and prefixes of J genes. The inference step derives D genes from the set of trimmed CDR3s and combines them with the set of known D genes (if available). The final step computes usage of D genes (including analysis of the allele usage of heterozygous D genes) and finds CDR3s with tandem D-D fusions.

Analysis of simulated CDR3s suggests that IgScout correctly reconstructs long D genes (length at least 20 nucleotides) if they give rise to at least 1% of CDR3s but misses short D genes (length <20 nt) if they give rise to <2.5% of CDR3s (see Supplemental Note “Benchmarking IgScout on simulated immunosequencing datasets”).

Since it is difficult to distinguish amplification artifacts from SHMs, IgScout takes a conservative approach and partially removes the clonal diversity (step “Hamming Graph (HG) Constructor” in Figure 3) to avoid propagation of amplification errors. Since naïve B cells do not have SHMs, the preprocessing step results in correcting amplification errors and enables reconstruction of long fragments of D genes. As a result, IgScout performs well on datasets with a sufficiently large number of consensus CDR3s (Figure 3). Below we analyze how the number of consensus CDR3s in real datasets affects the IgScout performance.

If a dataset contains hypermutated sequences, then the processing step keep SHMs in the consensus CDR3s. However, if the dataset does not have large clonal lineages (e.g., PBMC from a healthy donor) and the number of consensus CDR3 is large (Figure 3), IgScout treats unremoved SHMs as random errors and still reconstructs mutation-free D genes. However, if a dataset is formed by large clonal lineages, the preprocessing step creates a small number of consensus CDR3s with abundant



**FIGURE 1** | Abundances of all 443 common 15-mers (top) and the motif logo constructed for the most abundant 15-mer CGATTTTTGGAGTGG in the CDR3\* dataset constructed from the Set 1 dataset (bottom). (Top) The CDR3\* dataset contains 91% of all 15-mers appearing in human D genes (all 15-mers in human D genes are unique, i.e., appear in a single D gene). Four hundred forty-three common 15-mers in the CDR3\* set have abundances varying from 83 to 3,141. The y-axis represents the number of common 15-mers with given abundance (in logarithmic scale). Red, yellow, violet, and blue bars represent the number of common 15-mers with given abundance among known, mutated, trimmed, and foreign 15-mers, respectively. There exist 175 known, 195 mutated, 70 trimmed, and three foreign common 15-mers. The histogram represents 100 bins of width 30 each. (Bottom) The ATTTACGATTTTTGGAGTGGTTAT is the initial 28-nucleotide long sequence formed by positions in the motif logo with high information content (37). The motif logo was constructed using 3,141 sequences from the set CDR3\* containing the most abundant k-mer. After extending this 28-mer, IgScout reconstructed the 30-mer GTATTACGATTTTTGGAGTGGTTATTAT that is a substring of the 33-nucleotide long IGHD3-3 gene GTATTACGATTTTTGGAGTGGTTATTAT acc shown below the logo.

SHMs. Although IgScout is able to reconstruct some over-represented D genes for such datasets, some of the inferred D genes may still contain SHMs (Figure 3). We thus suggest to use caution while applying IgScout to clonally expanded datasets (see Supplemental Note “How IgScout results are affected by the number of consensus CDR3s and cell types”).

## Reconstruction of Human D Genes

IgScout is best suitable for reconstructing D genes in the case of naïve datasets and PBMC datasets with small clonal lineages. To illustrate this point, we applied IgScout to the NAÏVE, HEALTHY, ALLERGY, and HIV datasets. The number of consensus CDR3s in the NAÏVE datasets varies from 1,000 to 115,000. Figure 4 shows that IgScout reconstruct the same set of D genes as on the simulated datasets for naïve datasets



**FIGURE 2** | IgScout results on the *CDR3\** dataset. Each row shows a reconstructed string (strings are inferred in the order from the top to the bottom). Dark green segments correspond to reconstructed substrings of human D genes (flanking non-reconstructed nucleotides are shown in standard green). The most frequent 15-mers that were used for reconstructing the corresponding D genes are shown in red (their abundances are shown on the left). The reconstructed substring of the D2 gene (IGHD2-2) also occurs in D2\*2 and D2\*3 genes. Seventeen strings reconstructed by IgScout represent substrings of 17 human D genes. IgScout misses short prefixes and suffixes of D genes: 1.4 nucleotides on the left and 1.7 nucleotides on the right, on average for the Set 1 dataset (0.9 nucleotides on the left and 1.5 nucleotides on the right, on average after combining reconstructions over all HEALTHY datasets). IgScout did not reconstruct eight human D genes: D1 (IGHD1-1), D4 (IGHD4-4), D7 (IGHD1-7), D14 (IGHD1-14), D20 (IGHD1-20), D23 (IGHD4-23), D25 (IGHD6-25), and D27 (IGHD7-27) that contributed to few CDR3 in the Set 1. These genes have the following abundances of their most frequent 15-mers: 43 for D1, 59 for D4, 83 for D7, 0 for D14, 33 for D20, 75 for D23, 0 for D25, and 0 for D27.

with at least 20,000 consensus CDR3s. **Figure 4** shows that IgScout performs well on the HEALTHY and ALLERGY datasets and reconstructs the same set of D genes as for the simulated and NAÏVE datasets. Since number of consensus CDR3s in some of the HEALTHY and ALLERGY datasets is as low as 40,000, we recommend applying IgScout to dataset with small clonal lineages if the number consensus CDR3s exceeds 40,000. Although the HIV datasets also has many consensus CDR3s (varying from 19,000 to 55,000), the high SHM rate in the HIV datasets makes it difficult to reconstruct some short D genes (**Figure 4**). We thus suggest to use caution while applying IgScout to highly hypermutated datasets (such as repertoires of HIV and lymphoma patients).

**Figure 5** illustrates that IgScout reconstructed 18 out of 25 human D genes across all HEALTHY datasets, **Supplemental Note** “Summary of IgScout results across diverse immunosequencing datasets” describes inference of 20 human D genes across multiple immunosequencing datasets. **Supplemental Note** “Reconstructing variants of human D genes” describes inference of five allelic variants of the D7, D10, D16, D17, and D23 genes. However, since variations in D7, D17, and D23 genes affect the first or last nucleotides of the corresponding D genes, they likely represent computational artifacts caused by abundant nucleotides at the flanking positions of the D genes within CDR3s. In contrast, variations of the D10 and D16 genes

(referred to as D10+ and D16+, respectively) have mutations in the middle of D genes (**Figure 5**). They were inferred from multiple datasets (Set 5 and Set 7 for D10+, and Set 5, Set 7, Set 9, and Set 13 for D16+) and are consistent with alleles identified in previous studies [alleles IGH3-10\*p03 and IG3-16\*p03 reported in Lee et al. (41) and Boyd et al. (19)], but still missing in IMGT. **Supplemental Note** “Reconstructing variants of human D genes” illustrates that 50 (42) samples among 600 samples in the PROJECTS10 dataset support D10+ (D16+) variants and presents two more variants D10++ and D16++.

To demonstrate that D10+ and D16+ indeed represent new variants of D10 and D16 genes, we analyzed 40 whole genome sequencing datasets from the population-wide study of esophageal cancer (PRJNA427604 project) and searched for exact occurrences of D10+ and D16+ in reads. Both variations were detected in five out of 40 datasets (SRR6435661, SRR6435676, SRR6435686, SRR6435691, and SRR6435692) with the number of reads supporting D10+ (D16+) varying from 8 to 14 (30 to 58) across these five datasets.

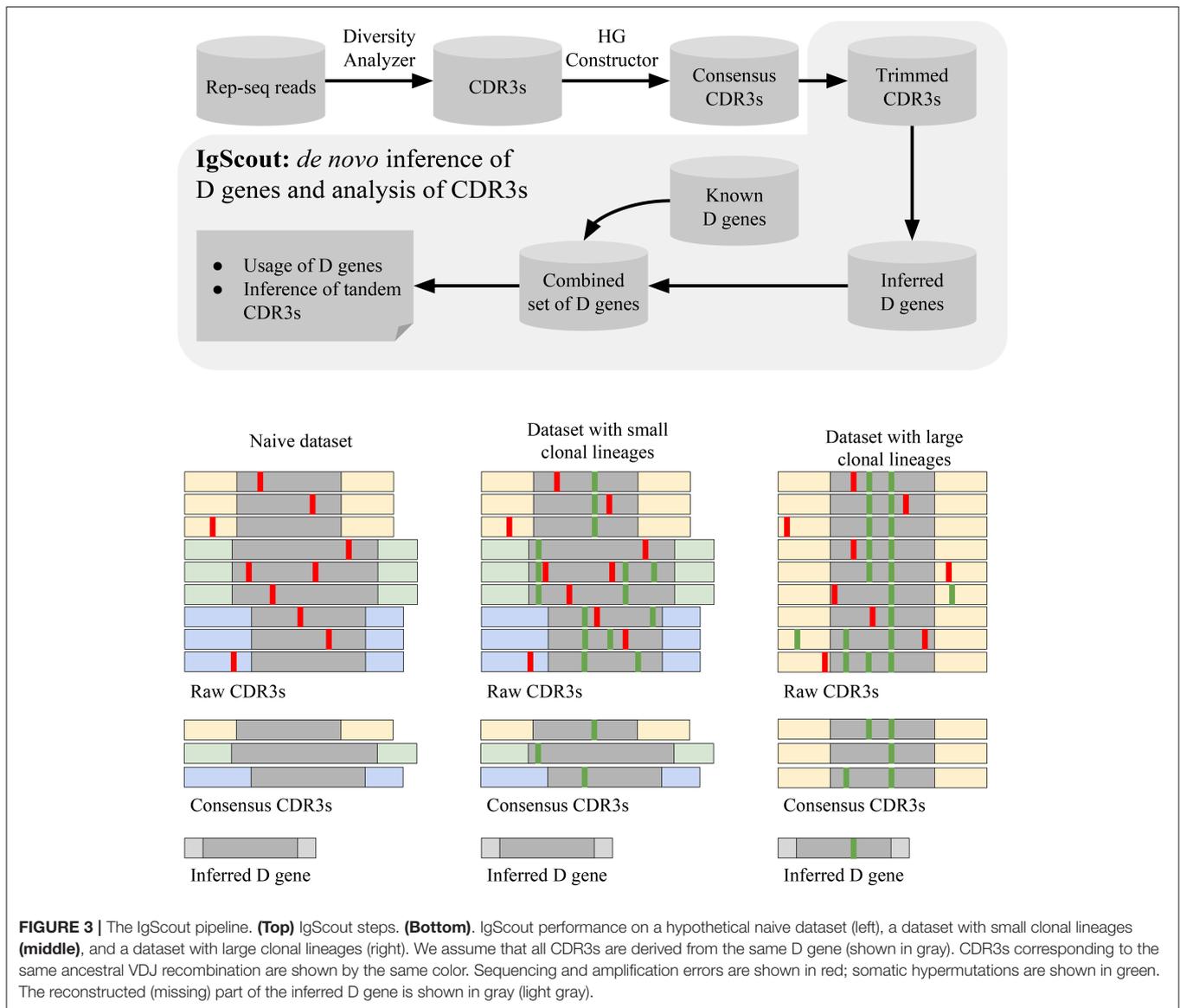
In general, IgScout has limitations with respect inferring both variants of a heterozygous D gene. Specifically, if two variants of the same D gene share a *k*-mer and IgScout selects this *k*-mer as a seed, the current version of IgScout may only reconstructs the most abundant variant of this D gene. We plan to enable inference of heterozygous D genes with two novel alleles and thus address this limitation in the next version of IgScout. Currently, to analyze allele usage of heterozygous human D genes, IgScout combines the inferred D genes with known D genes.

## Reconstruction of Camel D Genes

Although camel V genes were inferred in Conrath et al. (43), camel D genes remain unknown. We analyzed six CAMEL datasets from three camels (VH and VHH libraries for each camel) labeled as Camel 1VH, 1VHH, 2VH, 2VHH, 3VH, and 3VHH (33). While the VH libraries contain the heavy chain of the conventional (both heavy and light chain) camel antibodies, the VHH libraries contain the heavy chains of the *single-chain antibodies*.

We extracted camel CDR3s by aligning camel antibody repertoires against the known camel V and J genes using the IgReC tool (34). For the Camel 1VH dataset, IgScout constructed 60,066 consensus CDR3 sequences of average length 48 nucleotides. The *CDR3\** dataset for Camel 1VH has total length 1,400,360 nucleotides (the average length 23 nt).

IgScout reconstructed four D genes in the case of the Camel 1VH dataset that we refer to as D1, D2, D3, and D4 (see **Supplemental Note** “Reconstructing camel D genes”). It reconstructed four putative D genes in datasets Camel 1VHH, and Camel 2VH, and three putative D genes in the remaining three camel datasets (17 strings in total) that are largely consistent with genes D1, D2, D3, and D4 derived from the Camel 1VH dataset (previous studies assumed that the camel genome has a single germline D gene (43)). **Supplemental Note** “Reconstructing camel D genes” illustrates that all camel D genes are shared between the VH and VHH datasets. **Supplemental Note** “Usage of camel D genes” demonstrates that



the camel D genes have strikingly different usage in the VH and VHH antibodies.

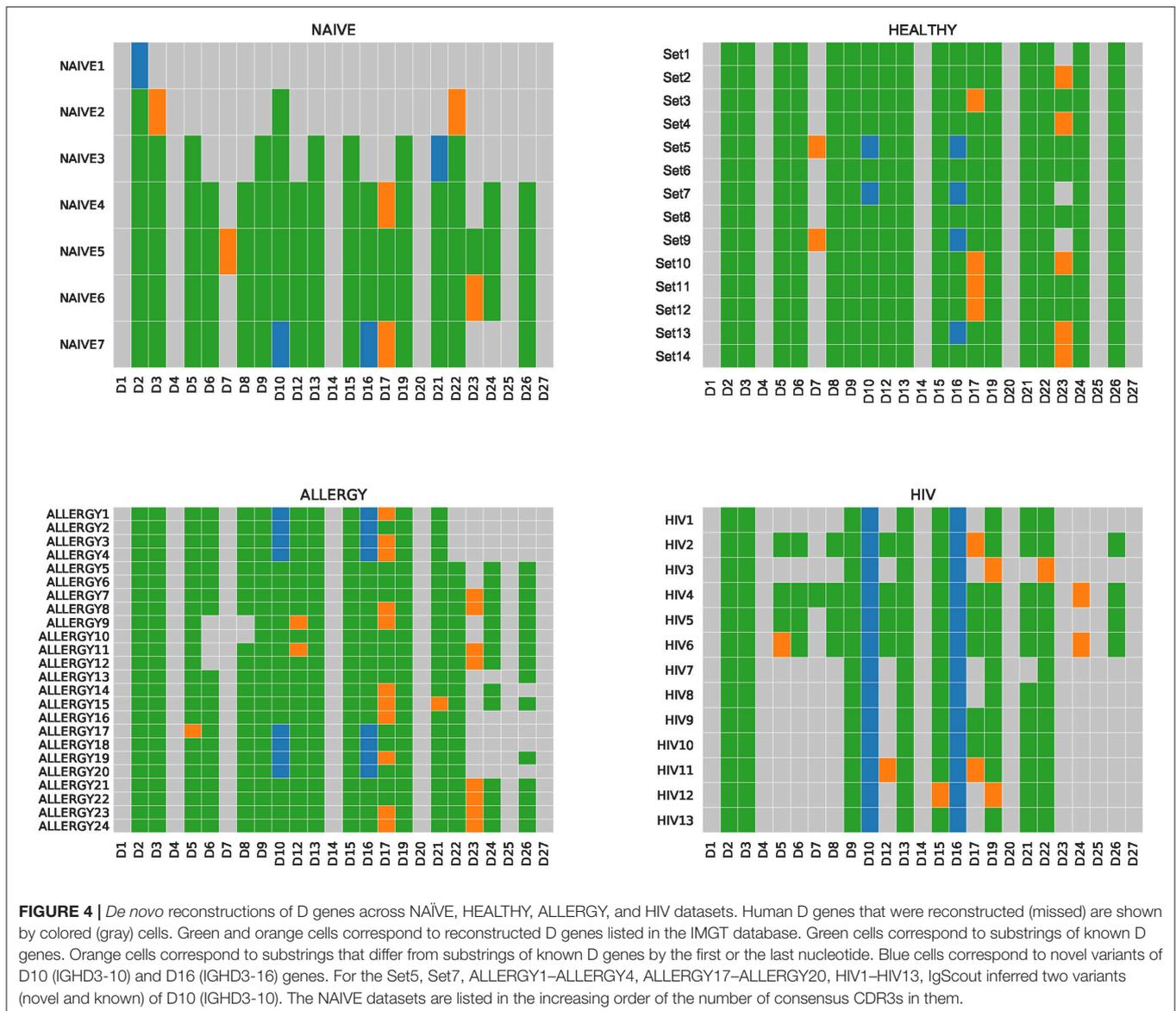
## D Gene Usage

Twenty-five human D genes form a set of strings that we refer to as *D-Genes*. Given an arbitrary string *Target*, a string *D* from *D-Genes*, and a parameter *k*, we say that a string *Target* is *formed* by *D* if it contains a *k*-mer from *D* but does not contain *k*-mers from other strings in *D-Genes* (the default value  $k = 11$ ). We classify a CDR3 as *traceable* if it is formed by a D gene and *non-traceable*, otherwise. The percentage of traceable CDR3s is rather conservative across all HEALTHY datasets:  $\approx 60\%$  of CDR3s in the HEALTHY datasets are traceable (**Supplemental Note** “Traceable CDR3s”).

Given a set of strings *Strings* and a string *D* from *D-Genes*, we define  $usage(Strings, D-Genes, D)$  as the fraction of traceable strings in *Strings* formed by the string *D*. We

are interested in  $usage(CDR3^*, D-Genes, D)$  for each human D gene. **Supplemental Note** “Traceable CDR3s” analyzes the usage of all human D genes across all HEALTHY datasets. **Supplemental Note** “D gene classification by IgScout and IgBlast” compares IgScout and IgBlast classification of D genes forming CDR3s.

We analyzed the usage of known and novel allelic variants ( $D10^+$  and  $D16^+$ ) across all HEALTHY datasets. **Figure 6** reveals that usage of allelic variants of D2 and D3 is consistent across all datasets with D2\*2 and D3 as dominant variants. However, the Set 5 has different dominant variants as compared to other datasets: D8\*2 (compared to D8 in all other datasets);  $D10^+$  (compared to D10 in all other datasets); and D21 (compared to D21\*2 in all other datasets). The variant  $D16^+$  is dominant in Sets 5, 7, 9, and 13, while the D16 gene is dominant in the remaining eight datasets.



## Tandem CDR3s

Given strings  $D$  and  $D'$ , and a parameter  $k$ , we say that a string *Target* is *formed* by  $D$  and  $D'$  if it contains  $k$ -mers from both  $D$  and  $D'$  and a  $k$ -mers from  $D'$  starts after a  $k$ -mer from  $D$  ends. Since tandem CDR3s represent a small fraction of all CDR3s, we set the default value  $k = 11$  (rather than  $k = 15$  for all CDR3s) to increase the number of identified tandem CDR3s. Although a smaller value of  $k$  may lead to identification of *pseudo-tandem* CDR3s, the Methods section describes how to filter out such pseudo-tandem CDR3s.

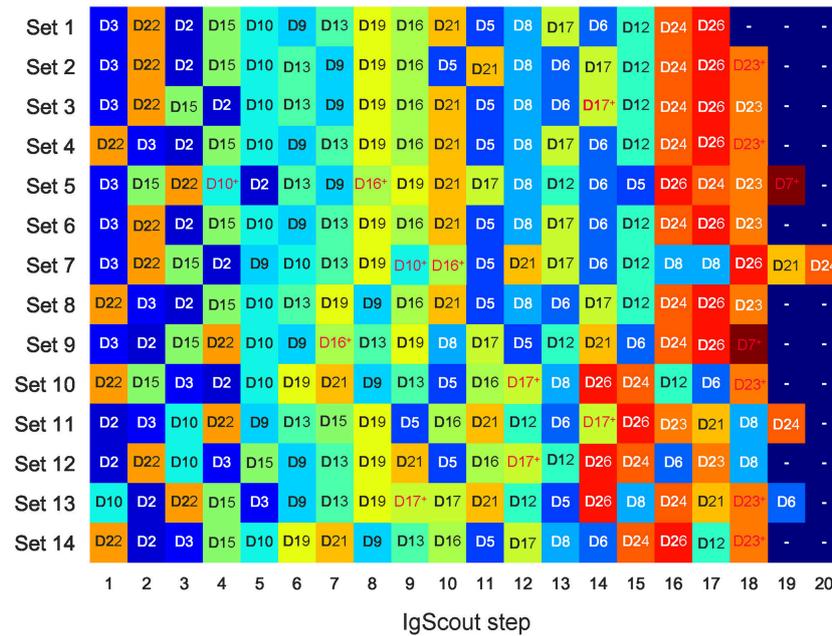
There exist 187 *tandem CDR3s* formed by two D genes in the *CDR3\** dataset (Figure 7). We denote the longest substring between a tandem CDR3 *Target* and  $D$  (*Target* and  $D'$ ) as  $D_{match}(D'_{match})$  and represent a tandem CDR3 *Target* as a concatenate of five strings *prefix* \*  $D_{match}$  \* *middle* \*  $D'_{match}$  \* *suffix*. We define the *span* of a tandem CDR3 formed by  $D$  and

$D'$  as the substring  $D_{match}$  \* *middle* \*  $D'_{match}$  and *inter-D* insertion as the substring *middle* (Figure 7).

Briney et al. (29) emphasized that detecting tandem CDR3s has to be done with caution since they are often confused with *pseudo-tandem* CDR3s formed by the standard V(D)J recombination (Figure 7). The Methods section describes how IgScout detects pseudo-tandem CDR3s. One hundred and fourteen out of 187 tandem CDR3s are not pseudo-tandem in the *CDR3\** dataset.

## Tandem Bias

There exists 114 tandem CDR3s in the Set 1 dataset and 1900 tandem CDR3s across all HEALTHY datasets. Figure 7 represents all tandem CDR3s as a *tandem matrix* and reveals that the vast majority of them correspond to cells in the upper half of this matrix. If tandem CDR3s were computational artifacts, we would expect similar numbers of CDR3s in the upper and lower



**FIGURE 5** | *De novo* reconstructions of D genes across HEALTHY datasets (**top**) and allelic variants D10<sup>+</sup> and D16<sup>+</sup> inferred by IgScout (**bottom**). (**Top**) Genes reconstructed at consecutive steps of IgScout for all HEALTHY datasets. Rows correspond to the datasets and columns correspond to the IgScout steps. Each cell is marked by a reconstructed D gene (each D gene is assigned a unique color). Cells marked with the “+” sign refer to strings that differ from known D genes by at most two nucleotides and correspond to putative novel variants (shown in red). (**Bottom**) Allelic variants D10<sup>+</sup> and D16<sup>+</sup> inferred by IgScout. Differences from human D genes and their allelic variants listed in the IMGT database are shown in red.

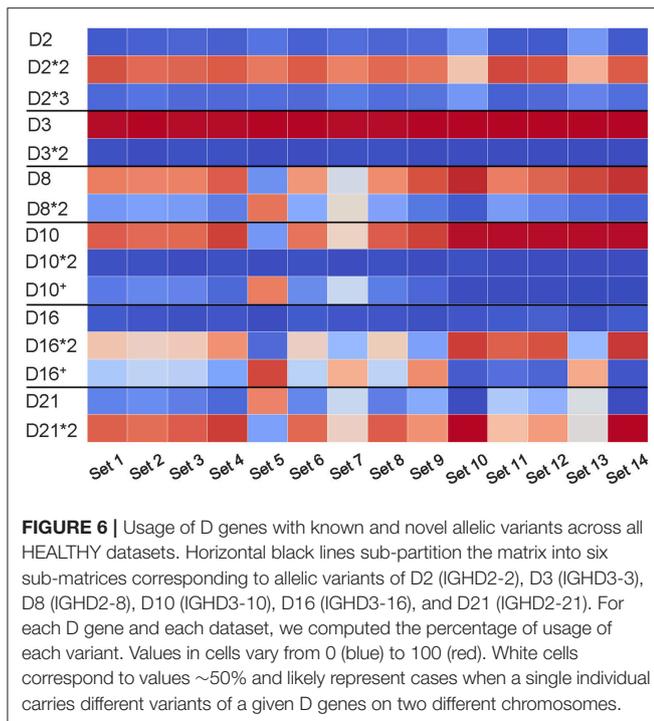
parts of the tandem matrix. We define the *tandem bias* as  $N_{lower} / (N_{upper} + N_{lower})$ , where  $N_{upper}$ , and  $N_{lower}$  is the sum of entries in the upper and lower parts of the tandem matrix, respectively (we assume that the main diagonal belongs to the lower part of the matrix). The tandem bias varies from 0.03 to 0.21% across various datasets (see **Supplemental Note**: “Analysis of tandem CDR3s”).

Since most pairs of D genes in tandem CDR3s contribute to the upper part of the tandem matrix (and thus follow the order of D genes in the IGH locus), entries in the lower part of the tandem matrix likely represents false positives. However, some of them may reveal possible duplications of D genes, e.g., the D22 row in the lower part of the tandem matrix in **Figure 7** reveals many tandem CDR3s. Analysis of the hepatitis patient 1,776 in the PROJECTS10 dataset (44) revealed particularly many entries in the D22 column in the lower part of the tandem matrix, suggesting a duplication of the D22 gene in this patient (see **Supplemental Note** “Analysis of tandem CDR3s”). Kidd et al. (23) analyzed biases in the

D-J pairing and also suggested that D22 may be duplicated in some individuals.

## Ultra-Long CDR3s Reveal Unusual Recombination Events

One thousand nine hundred tandem CDR3s across all HEALTHY datasets contain 1,081 distinct inter-D insertions, varying in length from 0 to 153 nucleotides. The two longest inter-D insertions (denoted  $I_1$  and  $I_2$ ) appear in the Set 1 and have length 153 nucleotides. They are formed by genes D9 and D10, differ by a single nucleotide, and appear in CDR3s differing by six nucleotides. Surprisingly, the inter-D insertion  $I_2$  coincides with the sequence of the IGH locus between the D9 and D10 genes. Germline D genes are flanked by *recombination signal sequences* (RSSs) with 12-nucleotide long spacer and the inter-D insertion  $I_2$  starts with the right RSS of D9 and ends with the left RSS of D10 (**Supplemental Note** “Ultra-long tandem CDR3s”).



Thus, ultra-long tandem CDR3s reveal unusual *RSS skipping* events during somatic recombination: skipping the right RSS of D9 and left RSS of D10 led to a tandem CDR3 representing a concatenate  $D9 + I_2 + D10$ . Although the found example is not productive, we also detected *RSS skipping* in nine productive ultra-long CDR3s across all HEALTHY and ALLERGY datasets. All productive CDR3s are formed by skipping of the right RSS of D22. Instead of it, somatic recombination uses a cryptic RSS (CACAGCA + ACCCAAACA) located at the distance 129 nt from the end of D22 and forms ultra-long CDR3s containing a genomic fragment of the IGH locus that starts with the right RSS of D22 (**Supplemental Note** “Ultra-long CDR3s”). The discovery of productive ultralong CDR3s challenges the conventional view of germline genes as non-overlapping substrings of DNA and reveals the first example of *nested* D genes, when one D gene is contained within another D gene.

The existing immunosequencing protocols are likely to miss ultra-long immunoglobulins since they are not designed to capture the abnormally long variable regions (exceeding ~400 nt). We captured reads containing ultra-long tandem CDR3s because the 300-nucleotide long paired reads (overlapping by only 50 nucleotides) in the Set 1 and ALLERGY datasets are longer than reads used in most other immunosequencing datasets. Thus, even if ultra-long tandem CDR3s were common, they would likely remain below the radar of most immunosequencing studies.

## Tandem CDR3s Contribute to Adaptive Immune Response

We investigated whether tandem CDR3s contribute to the adaptive immune response by analyzing their *isotypes*. Since IgG,

IgA, and IgE isotypes occur in plasma and memory B cells subjected to the antibody-antigen interactions, these isotypes they indicate (in difference from IgM isotypes common in memory and naïve B cells) that the corresponding antibodies participate in the adaptive immune response.

We inferred isotypes in the ALLERGY and HIV datasets using markers described in Levin et al. (31) (**Figure 8**). The vast majority of tandem CDR3s from the ALLERGY dataset correspond to the IgM isotype and thus are produced by memory and naïve B cells. In contrast, ~60% of tandem CDR3s in the HIV dataset correspond to the IgG type. This observation suggests that tandem CDR3s in the HIV-infected patients arise from immunoglobulins that are produced by plasma cells and thus might contribute to the immune response against HIV antigens.

## DISCUSSION

Since many human germline alleles remain unknown (particularly for non-European subjects), missing alleles may mislead clinical decisions (45) and lead to erroneous derivation of clonal lineages due to misinterpretations of SHMs. Thus, finding new germline alleles and building personalized sets of germline genes for each individual is important for downstream analysis of immunosequencing datasets.

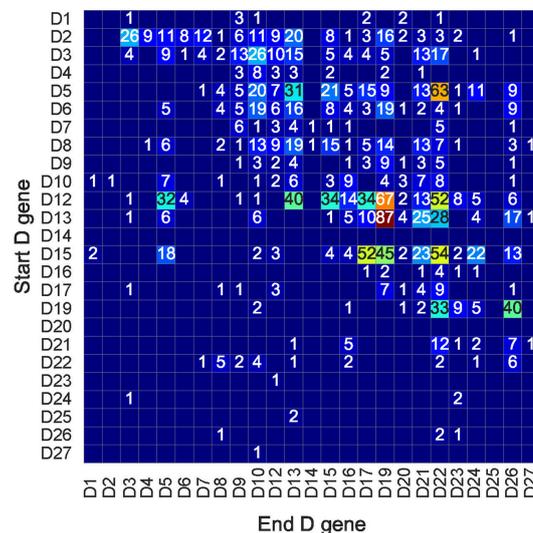
Although there exists a number of tools for inferring V and J genes (3, 21, 22), a more difficult problem of reconstructing D genes remains open. IgScout aims to reconstruct all D genes explaining a large percentage of the VDJ recombination in an antibody repertoire rather than to reconstruct all D genes. The IMGT database reflects the *genomic diversity* of D genes but not their *recombinant diversity* (information about rearrangements, transcription, and translation of D genes). Since assemblies of the highly repetitive IGH loci are fragmented and error-prone (7, 14, 42, 46) reconstruction of all germline genes from the whole-genome sequencing data is a difficult problem. Although the IGH locus is extremely diverse (16), it remains largely unknown how it varies across the human population. Moreover, even in the case when the IGH locus is correctly assembled, prediction of the functional germline genes is a non-trivial problem (2, 13).

Immunosequencing datasets reflect the recombinant diversity of antibody repertoires and thus complement the genomic datasets. If some D genes do not contribute to the VDJ recombination (e.g., our analysis suggests that genes D1, D14, D20, D25, and D27 do not significantly contribute to VDJ recombination in any of the analyzed datasets), they have limited contribution to immune response. In this paper, we focused on reconstructing D genes shaping the recombinant diversity rather than all D genes.

IgScout reconstructed 20 out of 25 human D genes across multiple datasets and missed genes D1, D14, D20, D25, D27 that form a small number of CDR3s (<0.1% each) across all analyzed datasets. It remains unclear whether some of these genes ever contribute to any CDR3s, for example genes D14 and D25 do not form any CDR3s in most datasets (few CDR3s formed by these D genes in some datasets may represent computational artifacts).

**D3:** GTATTACGATTTTGGAGTGGTTAtatacc  
**D5:** GTGGATACAGCTATGGttac  
**CDR3:** ACCACAGAACCCTTCAGTTTAGTCCGTATTACGATTTTGGAGTGGTTATCAGCCAGTGGATACAGCTATGGACCCGTTGACT  
**span:** GTATTACGATTTTGGAGTGGTTATCAGCCAGTGGATACAGCTATGG  
**inter-D insertion:** CAGCCA

**D10:** GTATTACTATGGTTCggggagtattataac  
**D16:** gtattatgattacgtttggGGGAGTTATGCTtatacc  
**CDR3:** GCGAGAGACAGTATTACTATGGTTCAGGGAGTTATGCGGGCTAACAACTACTACTACTACGGTATGGACGTC  
**span:** GTATTACTATGGTTCAGGGAGTTATGC  
**inter-D insertion:** A



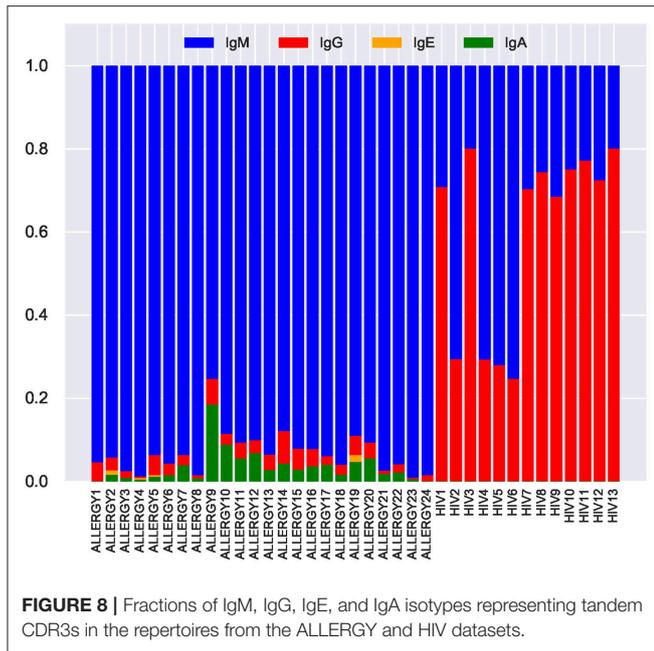
**FIGURE 7 |** A tandem CDR3 formed by genes D3 (IGHD3-3) and D5 (IGHD5-5) (**top**), a pseudo-tandem CDR3 formed by genes D10 (IGHD3-10), and D16 (IGHD3-16) (**middle**), and the tandem matrix for all tandem CDR3s across all HEALTHY datasets (**bottom**). (**Top**) A tandem CDR3 with  $D_{match}$ =GTATTAGGATTTTGGAGTGGTTAT,  $middle$ =CAGCCA, and  $D'_{match}$ =GTGGATACAGCTATGG. (**Middle**) The pseudo-tandem CDR3, formed by genes D10 (IGHD3-10) and D16 (IGHD3-16). This CDR3 was formed by a single gene D10 (IGHD3-10) with three mutations (shown in red). IgScout filters out most pseudo-tandem CDR3s. (**Bottom**) The number in a cell  $(i, j)$  shows the total number of tandem CDR3s formed by genes  $D_i$  and  $D_j$  across all HEALTHY datasets. Empty cells correspond to pairs of D genes that do not form tandem CDR3s. Genes D4 and D5 appear in two copies in the IGH loci. The second copy of D4 (IGHD4-11) appears between D10 (IGHD3-10) and D12 (IGHD5-12). The second copy of D5 (IGHD5-18) appears between D17 (IGHD4-17) and D19 (IGHD6-19). The vast majority of tandem CDR3 correspond to cells in the upper half of the matrix. The only populated column in the lower part of the tandem matrix corresponds to the D5 gene and likely results from tandem CDR3s formed by the second copy of D5 in the IGH locus.

IgScout revealed four new allelic variants ( $D10^+$ ,  $D10^{++}$ ,  $D16^+$ , and  $D16^{++}$ ), thus increasing the number of known variants of human D genes from 7 to 11. These new variants are unlikely to be computational artifacts since they were found in dozens immunosequencing datasets from distinct individuals and many whole genome sequencing datasets. The frequency of the already known Single Nucleotide Polymorphisms (SNPs) in D genes exceeds the frequency of SNPs in the entire human genome by two orders of magnitude (12 SNPs for all D genes of total length only 288 nucleotides).

Although IgScout revealed four novel variants of human D genes and inferred camel D genes, these genes will not be included in the IMGT database since they haven't been experimentally confirmed yet. Similarly to Gadala-Maria et al. (20), we argue that, like in other areas of genomics, the time has come to add such prediction to the IMGT database.

For example, the lion's share of genes in genomic databases represent computational predictions that haven never been experimentally confirmed. We argue that IMGT should classify alleles with varying levels of supporting evidence, not unlike classification systems used in other biological databases and in the recently established Open Germline Receptor Database (OGRDB), a new repository of germline genes maintained by The Adaptive Immune Receptor Repertoire (AIRR) Community (47).

Although IgScout is not specifically designed for reconstructing V and J genes, it turned out that its applications are not limited to reconstructing D genes (see **Supplemental Note** "De novo reconstruction of human J genes"). In addition to de novo reconstruction of D genes, it also detects tandem CDR3s. Briney et al. (29) postulated that tandem CDR3s mostly appear in naïve B cells and thus do not contribute



to adaptive immune response. In contrast, our analysis revealed that  $\sim 60\%$  of tandem CDR3s in the HIV dataset correspond to plasma and memory B cells.

## METHODS

### Inferring Germline Genes as the Trace Reconstruction Problem

In information theory, a string  $S$  yields a collection of *traces*, where each trace is independently obtained from  $S$  by substituting each symbol in  $S$  by another symbol from a fixed alphabet with a given probability  $\delta$ . Given the traces and the value  $\delta$ , the *Trace Reconstruction Problem* (35) is to reconstruct the original string  $S$ . *De novo* reconstruction of D genes results in a more complex version of the Trace Reconstruction Problem where traces are generated by multiple strings and each trace is obtained from one of these strings by (i) randomly trimming it from both sides, (ii) adding a randomly generated prefix in the front of the string, and (iii) adding a randomly generated suffix in the end of the string. Given a set of such traces (modeled by a set of trimmed CDR3s extracted from an immunosequencing dataset), the goal is to reconstruct the original set of strings.

### Extending the Consensus String

IgScout trims all positions of the motif logo with the information content below  $IC$  and computes the consensus string. Afterwards, it extracts the first  $k$ -mer of the consensus string and finds all CDR3s that contain this  $k$ -mer. If the position preceding the

## REFERENCES

1. 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 Protocols*. (2016) 11:1599–616. doi: 10.1038/nprot.2016.093

first  $k$ -mer in these reads has information content exceeding a threshold, IgScout adds the most frequent nucleotide at this position to the consensus and iterates. Afterwards, it applies a similar procedure to the last  $k$ -mer of the consensus string. The resulting extended consensus is reported as a putative D gene (Figure 1).

### Detecting Pseudo-Tandem CDR3s

Given strings  $Span$  and  $S$ , we define  $distance_t(Span, Target)$  as the minimum Hamming distance between  $t$ -mers in  $Span$  and  $S$ . Given a parameter  $\Delta$  (the default value  $\Delta = 5$ ) we define the  $\Delta$ -distance between strings  $Span$  and  $Target$  as  $distance_t(S, Target)$  for  $t = |Span| - \Delta$ , where  $|Span|$  stands for the length of the string  $Span$ . Finally, we define the  $\Delta$ -distance between a string  $Span$  and a set of strings  $Strings$  as the minimum  $\Delta$ -distance between  $Span$  and all strings in  $Strings$ .

We computed the  $\Delta$ -distance between the spans of all 187 identified tandem CDR3s in  $CDR3^*$  and all string in  $D$ -Genes. Seventy-three out of these 187 CDR3s can be explained as CDR3s originating from a single D gene (for the  $\Delta$ -distance threshold three). However, the remaining 114 CDR3s have  $\Delta$ -distance at least nine. We thus classify a CDR3 sequence  $Target$  formed by genes  $D$  and  $D'$  as pseudo-tandem if the  $\Delta$ -distance between the span of this pseudo-tandem CDR3 and  $D$ -Genes does not exceed a predefined threshold (the default value is three), and (truly) tandem, otherwise. See **Supplementary Note** “List of tandem CDR3s.”

## AUTHOR CONTRIBUTIONS

YS implemented the IgScout algorithm and performed benchmarking. YS and PP conceived the study, developed the IgScout algorithm, designed the computational experiments, and wrote the manuscript.

## FUNDING

YS was supported by the Data Science Fellowships at UCSD. The work of PP was supported by the NIH 2-P41-GM103484PP grant.

## ACKNOWLEDGMENTS

Authors are grateful to Dmitry Chudakov for providing us with the datasets Set 1–Set 9.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2019.00987/full#supplementary-material>

2. Wang Y, Jackson KJ, Sewell WA, Collins A. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol*. (2008) 86:111–5. doi: 10.1038/sj.icb.7100144
3. Ralph DK, Matsen FA IV. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput Biol*. (2016) 12:e1004409. doi: 10.1371/journal.pcbi.1004409

4. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res.* (2012) 40:e134. doi: 10.1093/nar/gks457
5. McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen FA IV. Quantifying evolutionary constraints on B-cell affinity maturation. *R Soc.* (2015) 370:20140244. doi: 10.1098/rstb.2014.0244
6. 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) 197:3566–3574. doi: 10.4049/jimmunol.1502263
7. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.* (2012) 13:363–73. doi: 10.1038/gene.2012.12
8. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe.* (2013) 13:691–700. doi: 10.1016/j.chom.2013.05.008
9. Chang CJ, Chen CH, Chen BM, Su YC, Chen YT, Hershfield M, et al. A genome-wide association study identifies a novel susceptibility locus for the immunogenicity of polyethylene glycol. *Nat Commun.* (2012) 8:522. doi: 10.1038/s41467-017-00622-4
10. Boyd SD, Liu Y, Wang C, Martin V, Dunn-Walters DK. Human lymphocyte repertoires in ageing. *Curr Opin Immunol.* (2013) 25:511–515. doi: 10.1016/j.coi.2013.07.007
11. 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) 188:1333–40. doi: 10.4049/jimmunol.1102097
12. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett A, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep.* (2016) 6:20842. doi: 10.1038/srep23876
13. Collins AM, Wang Y, Roskin KM, Marquis C, Jackson KJ. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos Trans R Soc Lond Ser B Biol Sci.* (2010) 370:20140236. doi: 10.1098/rstb.2014.0236
14. Luo S, Yu JA, Li H, Song YS. World-wide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *Life Sci Alliance.* (2019) 2:e201800221. doi: 10.26508/lsa.201800221
15. Yu Y, Ceredig R, Seoighe C. A database of human immune receptor alleles recovered from population sequencing data. *J Immunol.* (2017) 198:3758. doi: 10.4049/jimmunol.1700301
16. Watson CT, Matsen FA IV, Jackson K. JL, 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) 198:3371–3. doi: 10.4049/jimmunol.1700306
17. Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR. Automated de novo protein sequencing of monoclonal antibodies. *Nat Biotech.* (2008) 26:1336–8. doi: 10.1038/nbt1208-1336
18. Castellana NE, Pham V, Arnott D, Lill JR, Bafna V. Template proteogenomics: sequencing whole proteins using an imperfect database. *Mol Cell Proteomics.* (2010) 9:1260–70. doi: 10.1074/mcp.M900504-MCP200
19. 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) 184:6986–6992. doi: 10.4049/jimmunol.1000445
20. 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. *PNAS USA.* (2015) 112:E862–70. doi: 10.1073/pnas.1417683112
21. Corcoran MM, Phad GE, Vázquez Bernat N, 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. doi: 10.1038/ncomms13642
22. 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. doi: 10.3389/fimmu.2016.00457
23. Kidd MJ, Jackson KJ, Boyd SD, Collins AM. DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. *J Immunol.* (2016) 196:1158–64. doi: 10.4049/jimmunol.1501401
24. Kirik U, Greiff L, Levander F, Ohlin M. Data on haplotype-supported immunoglobulin germline gene inference. *Data Brief.* (2017) 13:620–40. doi: 10.1016/j.dib.2017.06.031
25. Meek KD, Hasemann CA, Carpa DJ. Novel rearrangements at the immunoglobulin D locus. Inversions and fusions add to IgH somatic diversity. *J Exp Med.* (1989) 170:39–57. doi: 10.1084/jem.170.1.39
26. Kurosawa Y, Tonegawa S. Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. *J Exp Med.* (1982) 155:201–8. doi: 10.1084/jem.155.1.201
27. Corbett SJ, Tomlinson IM, Sonnhammer EL, Buck D, Winter G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, “minor” D segments or D-D recombination. *J Mol Biol.* (1997) 270:587–97. doi: 10.1006/jmbi.1997.1141
28. Watson LC, Moffatt-Blue CS, McDonald RZ, Kompfner E, Ait-Azzouzen D, Nemazee D, et al. Paucity of V-D-D-J rearrangements and VH replacement events in lupus prone and nonautoimmune TdT<sup>-/-</sup> and TdT<sup>+/+</sup> mice. *J Immunol.* (2006) 177:1120–8. doi: 10.4049/jimmunol.177.2.1120
29. Briney BS, Willis JR, Hicar MD, Thomas JW, Crowe JE. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology.* (2012) 137:56–64. doi: 10.1111/j.1365-2567.2012.03605.x
30. Larimore K, McCormick MW, Robins HS, Greenberg P. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol.* (2012) 189:3221–30. doi: 10.4049/jimmunol.1201303
31. 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) 139:1026–30. doi: 10.1016/j.jaci.2016.06.040
32. Landais E, Murrell B, Briney B, Murrell S, Rantalainen K, Berndsen Z, et al. HIV envelope glycoform heterogeneity and localized diversity govern the initiation and maturation of a V2 apex broadly neutralizing antibody lineage. *Immunity.* (2017) 47:990–1003.e9. doi: 10.1016/j.immuni.2017.11.002
33. 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:e0161801. doi: 10.1371/journal.pone.0161801
34. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, Pevzner PA. Reconstructing antibody repertoires from error-prone immunosequencing reads. *J Immunol.* (2017) 199:3369–80. doi: 10.4049/jimmunol.1700485
35. Mitzenmacher MA. Survey of results for deletion channels and related synchronization channels. *Prob Survays.* (2009) 6:1–33. doi: 10.1214/08-PS141
36. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics.* (2005) 11:351–8. doi: 10.1093/bioinformatics/bti1018
37. Crooks GE, Hon G, Chandonia J-M, Brenner S. WebLogo: a sequence logo generator. *Genome Res.* (2004) 14:1188–90. doi: 10.1101/gr.849004
38. Thörnqvist L, Ohlin M. Critical steps for computational inference of the 3'-end of novel alleles of immunoglobulin heavy chain variable genes - illustrated by an allele of IGHV3-7. *Mol Immunol.* (2018) 103:1–6. doi: 10.1016/j.molimm.2018.08.018
39. Thörnqvist L, Ohlin M. Data on the nucleotide composition of the first codons encoding the complementary determining region 3 (CDR3) in immunoglobulin heavy chains. *Data Brief.* (2018) 19:337–52. doi: 10.1016/j.dib.2018.04.125
40. Thörnqvist L, Ohlin M. The functional 3'-end of immunoglobulin heavy chain variable (IGHV) genes. *Mol Immunol.* (2018) 96:61–68. doi: 10.1016/j.molimm.2018.02.013
41. Lee CE, Gaëta B, Malming HR, Bain M, Sewell WA, Collins AM. Reconsidering the human immunoglobulin heavy-chain locus: 1. An evaluation of the expressed human IGHD gene repertoire. *Immunogenetics.* (2006) 57:917–925. doi: 10.1007/s00251-005-0062-5

42. Matsuda F, Ishii K, Bourvagnet P, Kuma KI, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med.* (1998). 188:2151–62. doi: 10.1084/jem.188.11.2151
43. Conrath K, Wernery U, Muyldermans S, Nguyen V. Emergence and evolution of functional heavy-chain antibodies in Camelidae. *Dev Comp Immunol.* (2003) 27:87–103. doi: 10.1016/S0145-305X(02)00071-X
44. Galson JD, Trück J, Clutterbuck EA, Fowler A, Cerundolo V, Pollard AJ, et al. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med.* (2016) 8:68. doi: 10.1186/s13073-016-0322-z
45. Xochelli A, Agathangelidis A, Kavakiotis I, Minga E, Sutton LA, Baliakas P, et al. Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukemia. *Immunogenetics.* (2015) 67:61–66. doi: 10.1007/s00251-014-0812-3
46. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet.* (2013) 92:530–46. doi: 10.1016/j.ajhg.2013.03.004
47. Ohlin M, Scheepers C, Corcoran M, Lees William D, Busse Christian E, Davide B, et al. Documentation, and naming. *Front Immunol.* (2019) 10:435. doi: 10.3389/fimmu.2019.00435

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Safonova and Pevzner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.