Omer *et al. Genome Medicine* (2022) 14:2
https://doi.org/10.1186/s13073-021-01008-4

Genome Medicine

RESEARCH Open Access

T cell receptor beta germline variability is revealed by inference from repertoire data

Aviv Omer[1,2†], Ayelet Peres[1,2†], Oscar L Rodriguez[3], Corey T Watson[3], William Lees[4], Pazit Polak[1,2], Andrew M Collins[5] and Gur Yaari[1,2*]

Abstract

Background: T and B cell receptor (TCR, BCR) repertoires constitute the foundation of adaptive immunity. Adaptive immune receptor repertoire sequencing (AIRR-seq) is a common approach to study immune system dynamics. Understanding the genetic factors influencing the composition and dynamics of these repertoires is of major scientific and clinical importance. The chromosomal loci encoding for the variable regions of TCRs and BCRs are challenging to decipher due to repetitive elements and undocumented structural variants.

Methods: To confront this challenge, AIRR-seq-based methods have recently been developed for B cells, enabling genotype and haplotype inference and discovery of undocumented alleles. However, this approach relies on complete coverage of the receptors' variable regions, whereas most T cell studies sequence a small fraction of that region. Here, we adapted a B cell pipeline for undocumented alleles, genotype, and haplotype inference for full and partial AIRR-seq TCR data sets. The pipeline also deals with gene assignment ambiguities, which is especially important in the analysis of data sets of partial sequences.

Results: From the full and partial AIRR-seq TCR data sets, we identified 39 undocumented polymorphisms in T cell receptor Beta V (TRBV) and 31 undocumented 5′ UTR sequences. A subset of these inferences was also observed using independent genomic approaches. We found that a single nucleotide polymorphism differentiating between the two documented T cell receptor Beta D2 (TRBD2) alleles is strongly associated with dramatic changes in the expressed repertoire.

Conclusions: We reveal a rich picture of germline variability and demonstrate how a single nucleotide polymorphism dramatically affects the composition of the whole repertoire. Our findings provide a basis for annotation of TCR repertoires for future basic and clinical studies.

Keywords: AIRR-seq, Genotype, Allele inference, TRB, TCR, Immune repertoires

*Correspondence: gur.yaari@biu.ac.il
† Aviv Omer and Ayelet Peres contributed equally to this work.
[1] Faculty of Engineering, Bar Ilan University, 5290002 Ramat Gan, Israel
[2] Bar Ilan institute of Nanotechnology and Advanced Materials, Bar Ilan University, 5290002 Ramat Gan, Israel
Full list of author information is available at the end of the article

Background

The immune system's success in fighting countless evolving pathogens depends on a dynamic and diverse set of B and T cell receptors. Due to the longevity of immunological memory, high-throughput sequencing of adaptive immune receptor repertoires (AIRR-seq) provides detailed insights into the past and present encounters of the human immune system [1]. It can teach us about fundamental immune processes and reveal dysregulation, with broad implications for biomedicine. B and T cell receptors are assembled within B and T cells, respectively, during differentiation from hematopoietic stem cells, by a complex process involving somatic recombination of a large number of germline-encoded Variable (V), Diversity (D), and Joining (J) gene segments, along with junctional diversity in the form of addition and subtraction of nucleotides at the boundaries where these segments are joined together [2]. This V(D)J recombination process creates a diverse repertoire of receptors that together with the innate immune system form the first line of defense against pathogens.

Genetic factors are expected to influence the structure and functionality of AIRRs [3, 4]. However, understanding these genetic effects is confounded by lack of knowledge about the population genetics of the TCR and BCR encoding genomic loci, and the special challenges involved in describing the germline gene set of any individual [5]. The lack of knowledge about these loci is due to difficulties in reliably mapping repetitive elements and undocumented structural variations with short-read sequencing. Many TCR genes were reported in the decade after their first discovery [6]. Complete sequences of the TCR encoding loci were reported in 1985 [7–9], and this led to the development of the TCR encoding gene nomenclature by the ImMunoGeneTics (IMGT) group [10]. Since then, very few allelic variants have been entered into the IMGT reference directories of germline genes. In fact, no new allelic variants of the TCR variable region genes have been named this century despite published studies suggesting that the IMGT TCR reference directory (www.imgt.org/download/V-QUEST/) may be far from complete [11–13].

Until recently, there has been a similar lack of attention paid to the documentation of BCR encoding loci, because the direct genomic sequencing of these loci is also very challenging [14, 15]. This changed with the development of a method for targeted long-read direct sequencing of the BCR heavy chain encoding locus [16], and with the realization that BCR encoding undocumented germline alleles and genotypes can be reliably inferred from AIRR-seq data [17–20], as well as haplotypes [21, 22], and chromosomal deletions within the BCR encoding loci [23]. Even though TCR V(D)J gene rearrangements are generated by analogous mechanisms to BCR rearrangements, to date, there is no published data about TCR germline allele inference and structural variation in the TCR encoding loci. Recently, there were attempts to extract BCR and TCR encoding allelic information from short-read whole-genome sequencing data [11, 24, 25], but these approaches were not validated with targeted sequencing or AIRR-seq and therefore are subject to criticism regarding the reliability of the inferences [5, 26]. Hence, to study genomic variations in TCR encoding loci and their relations to the expressed repertoires, there is a need to adapt BCR inference tools to TCR data.

In T cells, due to the lack of somatic mutations, most studies sequence only a small fraction of the variable region. AIRR-seq data can be generated by methods that differ in the length of their coverage of V(D)J sequences. 5′ RACE amplifies the whole V(D)J region from the 3′ end of the J region to the 5′ end of the mRNA molecule. BIOMED-2 primers [27] amplify partial V(D)J sequences from the J gene to the framework-2 (FR2) of the V gene, while the Adaptive Biotechnologies [28] approach generates only 87 nucleotides from a fixed position within each T cell receptor Beta J (TRBJ) gene in FR4 and includes the complementary determining region 3 (CDR3) and a fraction of the TRBV gene from FR3. As there are TRBV alleles that are identical to other TRBV alleles for varying lengths from their 3′ ends, in some cases, it is impossible to identify the gene and allele source of these partial V(D)J sequences, generating a serious gene assignment ambiguity problem. Thus, the development of a TRBV genotype inference method requires the detailed documentation of any gene pairs that can be impossible to distinguish for a given sequencing approach.

Here, we adapt a B cell inference pipeline to TRB AIRR-seq data (Fig. 1A), to work with these three types of sequencing approaches (Fig. 1B), and infer undocumented alleles, i.e., alleles that were previously not documented in IMGT, as well as single and double chromosome deletions, genotypes, and haplotypes. The pipeline was adapted to deal with the gene assignment ambiguity problem, which is especially important in the analysis of data sets of partial sequences (Fig. 1C, D). We applied this pipeline to four of the largest AIRR-seq data sets currently available and revealed a rich picture of germline variability and a demonstration of how a single nucleotide polymorphism dramatically affects the composition of the whole repertoire. TCR germline variability and its effects on the expressed repertoire may have important implications for TCR-based immunotherapy and disease diagnosis, for example by enabling the discovery of TCR-related predispositions to diseases, or responses to different therapies.

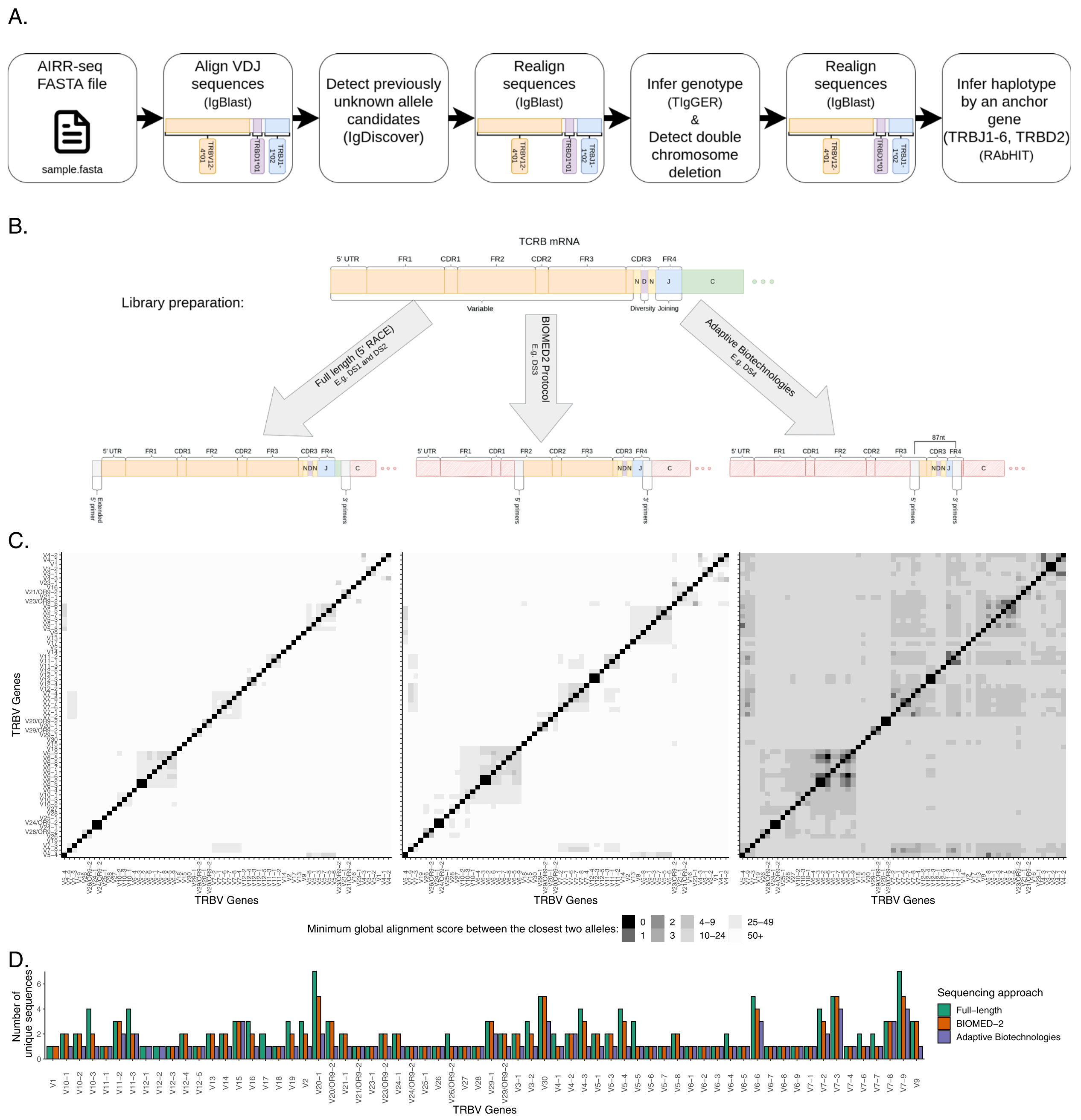

Fig. 1 Overview of TRB genotyping and haplotyping workflow, TCRβ sequencing technologies, and number of documented alleles. **A** An illustration of the steps for TRB genotyping and haplotyping. The input is an AIRR-seq fasta file after preprocessing, according to the sequencing protocol described in the "Methods" section (see the "Data" section). **B** An illustration of the three TCRβ sequencing approaches investigated in this study. **C** Heatmaps of distances between amplified TRBV gene sequences. The three panels correspond to the sequencing approaches shown above them. Colors correspond to the distances between the genes as measured by the minimum global alignment score (with a padded beginning and a penalty of one for each mismatch, insertion, or deletion) between the closest two alleles in the sequenced region. **D** Barplot of the unique sequence number per gene in the studied sequencing approaches

Methods

Data

Four AIRR-seq TCR data sets (DSs) were collected [29–34] (Table 1). Three were bulk sequenced, but each of them was generated using a different sequencing protocol, producing sequences of different lengths. The DSs also differ in sequencing depth, i.e., the number of unique sequences per repertoire (Additional file 1: Fig. S1). Another DS of TCRs amplified from single cells was collected from three different sources [33–35] (see Additional file 1: Table S1). The data sets are described in Table 1. In DS3, there were originally 313 individuals and 348 samples, out of which 108 samples were obtained from 94 individuals with hematological cancer. These 108 samples were dropped out of the analysis, given their likely bias towards substantial mutation and oligoclonality. DS2 and DS4 were downloaded after preprocessing, DS1 was preprocessed according to the preprocessing of Eliyahu et al. [29], and DS3 was preprocessed using pRESTO [36] according to the example workflow "Illumina MiSeq 2x250 BCR mRNA" as follows: (i) paired ends were assembled, (ii) sequences with low quality (mean Phred quality scores lower than 20) were removed, (iii) the 3′ and 5′ end primers were cut, and (iv) duplicate sequences were removed and collapsed.

Merging indistinguishable genes

Sequences of two full-length TRBV genes, TRBV6-2*01 and TRBV6-3*01, are indistinguishable (Fig. 1C). We therefore refer to them here as TRBV6-2*01/TRBV6-3*01. TRBV sequences amplified using the BIOMED-2 primers are partial, yet it is still possible to differentiate most of the genes. Only TRBV6-2 and TRBV6-3, as well as TRBV12-3 and TRBV12-4, could not be differentiated (Fig. 1C). Those indistinguishable sequences are referred to here as TRBV6-2/TRBV6-3 and TRBV12-3/TRBV12-4. The Adaptive Biotechnologies sequencing protocol generates very short partial TRBV gene sequences, yet it is still possible to identify most of them. Only the gene pairs TRBV6-2/TRBV6-3, TRBV12-3/TRBV12-4, TRBV3-1/TRBV3-2, and TRBV6-5/TRBV6-6 were indistinguishable (Fig. 1C).

Adaptive Biotechnologies supplies 87 nucleotides from a fixed position within each TRBJ gene [28]. As a result, the given coverage of the TRBV segment is not constant, because TRBJ genes and junction regions have different lengths. Therefore, the distribution of the first position that the sequences covered of the TRBV reference was investigated. Ninety-six percent of the sequences cover the first position following the TRBV gene primers. The BIOMED-2 protocol did not include primers for TRBV12-2. Thus, we were unable to explore the usage or genetic variation of this gene in DS3.

Allele pattern collapsing

Although there are few ambiguities in the identification of partial TRBV genes, the unambiguous identification of partial allelic variants is more problematic. Many SNPs that distinguish between alleles are located outside the regions that are generated using BIOMED-2 or Adaptive Biotechnologies primers. Thus, all alleles were collapsed into partial allelic variation groups. The sequence of each partial allelic variation group was determined to be identical to the longest allele sequence reference (out of the identical partial alleles' references). The allele patterns were named here using the following structure: [gene name]*[protocol primers][0-9][0-9]. The BIOMED-2 partial allelic variants were symbolized by bp, and the Adaptive Biotechnologies partial allelic variants were symbolized by ap. For example, the partial sequence of the allele TRBV5-6*01 was collapsed into the partial allelic variation groups TRBV5-6*bp01 and TRBV5-6*ap01 (see Additional file 1: Table S2 and Additional file 1: Table S3).

Table 1 AIRR-seq TCR data sets (DS) analyzed in this study

Data set	Cohort	# of individuals	# of samples	Sequencing protocol	UMI	Helix	Accession	Citation
DS1	Hepatitis C Virus (HCV)	28	28	5′ RACE (full-length)	+	RNA	ENA:PRJEB28370	Eliyahu et al. [29]
DS2	-	25	25	10x Genomics (full-length)	+	RNA	https://www.10xgenomics.com/resources/datasets	10x Genomics [32]
							EGA:EGAS00001003449	Wen et al. [33]
							GEO:GSE145926	Liao et al. [34]
DS3	Cancer	219	240	BIOMED-2	-	DNA	ENA:PRJEB33490	Simnica et al. [30]
DS4	Cytomegalovirus (CMV)	786	786	Adaptive Biotechnologies	-	DNA	https://doi.org/10.21417/B7001Z	Emerson et al. [31]

The primers of the BIOMED-2 protocol were taken from van Dongen et al. [27], and the primers of Adaptive Biotechnologies were taken from Robins et al. [28].

Genotype and undocumented allele inference

IgDiscover [17] was used for detection of undocumented allele candidates, and TIgGER for genotype inferences. Sequences were first aligned with IgBlast and processed using the IgDiscover igblast function and the Change-O MakeDb function [37]. A correction to the inferred CDR3 sequences in the IgDiscover output was done by replacing the sequences with their counterparts in the MakeDb output. TRBV allele candidates were inferred using IgDiscover's "discover" function. Undocumented allele candidates were filtered based on several rules. First, suspected SNPs were counted only between the boundaries: 5′ position of N+5 where N is the nucleotide position in which the primer ends or the sequence starts (the larger between the two), and 3′ position 316 by IMGT numbering. Second, candidate undocumented alleles were filtered out if they were not an exact match to at least 5% of the gene alignments. Third, such candidates had to have a sufficient rearrangement diversity: at least two different CDR3 lengths and two TRBJ genes. Fourth, for noisy data sets in which chimeras were observed, candidates that could result from chimerism were filtered out. A candidate was suspected as having the potential to result from chimerism if two alleles from separate genes could generate an exact matched sequence in the range between nucleotides N+5 and 316.

Undocumented allele candidates were then combined with the IMGT TRBV Reference Directory to create individual-specific Reference Directories. Sequence sets were then re-aligned against the new directories, and genotypes were constructed with TIgGER using a Bayesian approach [19]. Genotyping was limited to sequences with a single assignment (only one best match) and with up to one mismatch in the TRBV segment. For the construction of the TRBD genotype, sequences with mismatches in the TRBD segment or with identifiable TRBD sequences shorter than nine nucleotides were filtered out. TIgGER's level of confidence was calculated using a Bayes factor (K) from the posterior probability for each model. The larger the K, the greater the certainty in the genotype inference. lk that is used throughout the manuscript indicates the log of K.

For the undocumented alleles, two additional filters after the genotype inferences were added. First, undocumented alleles were filtered out if there was more than one SNP within a stretch of four adjacent nucleotides [38]. Second, to account for potential sequencing errors, we investigated the modality of the usage distribution of each undocumented allele in the population. The alleles that did not follow the expected usage distribution of bi- or tri-modal were discarded.

Validating undocumented alleles/variants in long-read assemblies

Thirty-five diploid (70 haplotypes) whole-genome sequencing assemblies from 32 unrelated individuals and three offsprings [39] were downloaded and aligned to GRCh38 (Genome Reference Consortium Human Build 38) [40] using BLASR with default parameters [41]. Gene sequences (5′ UTR, leader-1, leader-2, and exons) from the assemblies were extracted based on the alignment. Undocumented alleles and variants were determined to be present in the assemblies only if they exactly matched the extracted gene sequence. Deleted genes were detected by their absence in the assembly and visually validated using IGV [42].

Determining the borders between TRBD2 genotype groups

We examined the fraction of TRBD2*01 assignments amongst all sequences unambiguously assigned to TRBD2 in DS1. The observed frequencies defined three distinct groups (Additional file 1: Fig. S2A). In the first group ($\sim 0 - 0.2$), TRBD2*01 assignments peaked around 0.125 of the TRBD2 annotations. The second group ($\sim 0.2 - 0.8$) peaked around 0.5, and the third group ($\sim 0.8 - 1$) peaked around 0.95. According to this distribution, individuals in the first group are homozygous for TRBD2*02, and all alignments to TRBD2*01 are in error. The second group corresponds to individuals who are heterozygous for TRBD2, and individuals in the third group are homozygous for TRBD2*01. The resulting genotype frequency distribution of TRBD2 were in Hardy–Weinberg equilibrium.

To better define the thresholds differentiating the three groups, we turned to the larger data sets. DS3 was too noisy for this purpose (Additional file 1: Fig. S2B), most likely due to the library preparation and sequencing protocol. In DS4, which contains 768 individuals, the TRBD2*01 distribution was tri-modal and very similar to that in DS1 (Additional file 1: Fig. S2C). The homozygous TRBD2*01 group is centered around a frequency of 0.96. The heterozygous group is centered around 0.45, and the homozygous TRBD2*02 group is centered around 0.12. The medians of the three groups are close to their means (Additional file 1: Table S4), and as the size of the groups is large enough, we used the central limit theorem and thus assume that the three data sets are normally distributed.

Boundaries between the tri-modal peaks were defined as the equilibrium points between the peaks. The equilibrium point between two normal distributions is the point between the two averages of the groups, at which the probability of being in either one of the two groups is equal. For two distributions $X_1 \sim N_1(\mu_1, \sigma_1)$

and $X_2 \sim N_2(\mu_2, \sigma_2)$, the equilibrium point x is determined as follows:

$$P_1(X_1 \geq x) = P_2(X_2 \leq x)$$

$$\Phi_1\left(-\frac{x-\mu_1}{\sigma_1}\right) = \Phi_2\left(\frac{x-\mu_2}{\sigma_2}\right)$$

$$\frac{\mu_1 - x}{\sigma_1} = \frac{x-\mu_2}{\sigma_2}$$

$$(\mu_1 - x)\sigma_2 = (\mu_2 - x)\sigma_1$$

$$\mu_1\sigma_2 - x\sigma_2 = \mu_2\sigma_1 - x\sigma_1$$

$$x\sigma_1 + x\sigma_2 = \mu_1\sigma_2 + \mu_2\sigma_1$$

$$x = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1 + \sigma_2}$$

Applying the above to DS4 resulted in the following borders: the homozygous TRBD2*02 group was composed of 192 samples with a fraction of TRBD2*01 lower than 0.2066, the heterozygous group was composed of 404 samples with a fraction between 0.2066 and 0.8968, and the homozygous TRBD2*01 group was composed of 190 samples with a TRBD2*01 fraction above 0.8968. Consistent with the Hardy-Weinberg principle, the genotype frequencies of TRBD2*01 homozygotes, TRBD2*02 homozygotes, and heterozygotes was 0.244, 0.242, and 0.514, respectively.

Gene usage comparison

The differences in gene usage between groups were analyzed using a two-tailed Mann–Whitney test with the p value for significance adjusted by the Bonferroni correction to deal with the problem of multiple comparisons.

Double chromosome deletion inference

The detection of double chromosome deletion was done using a published method [23] and adapted to TRB. Briefly, the method assesses whether a gene can be declared as deleted using a binomial test where the parameters for a given individual are as follows: X is set to the number of sequences mapped to a given gene, N is the total number of sequences, and P is the lowest relative frequency for the given gene from the non-deleted candidates. To determine a threshold for a given gene, a minimum cutoff is set and the closest frequency above the cutoff is set as (P). For TRB, the minimum cutoff of the average gene usage was lowered from 0.005 to 0.0005. A data frame table was used that contained the following columns: individual, gene, N, and total. N represents the number of unique sequences for each gene. The total column records the total number of the individual's unique sequences. A binomial test for detecting chromosome deletions was then applied to the data frame table.

Haplotype inference

RAbHIT was used as previously described [22], with TRBD2 and TRBJ1-6 anchors to infer TCR haplotypes. The epsilon error parameter was adjusted to deal with TRBD2 alignment errors and was estimated with reference to the frequency distribution of TRBD2*01 alignments amongst all TRBD2-bearing sequences (see Additional file 1: Fig. S2) mentioned above. As ~12.5% of TRBD2*02 rearrangements mis-align to TRBD2*01, the epsilon, a parameter which defines the probability of the mis-assignment, was set at 0.125 if TRBD2*02 dominated the TRBD2 alignments. Around 4% of the TRBD2*01 rearrangements mis-align to TRBD2*02, and epsilon was therefore set at 0.04 if TRBD2*01 dominated the alignments.

Results

Identification of undocumented alleles and upstream sequence variations within TRBV genotypes

To explore allelic variation in the TRBV locus, we inferred the sets of alleles carried for each expressed TRBV gene in many individuals, i.e., personal genotypes. For this, we took a multi-step approach, using IgDiscover [17] and TIgGER [18], which yielded a set of candidate sequences that have not previously been documented in IMGT. Hereafter, we will refer to such sequences as "undocumented alleles." To ensure our confidence in the inference of undocumented alleles, we took steps that reduce the influence of sequencing errors (see the "Methods" section).

We applied the above approach to four data sets, spanning different sequencing protocols and scales (see the "Methods" section). From data set 1 (DS1), which includes 28 individuals sequenced in full length, we inferred 18 undocumented alleles (Fig. 2A). Four of the 18 alleles were only seen at very low levels — less than 20% of all identified alleles of those genes, and showed a unimodal distribution. This did not follow our expected usage distribution (these alleles are marked red in Fig. 2A). These candidate "alleles" therefore potentially result from sequencing errors. Two more undocumented alleles were considered erroneous due to the presence of more than one SNP in short nucleotide stretches. After discarding these six allele candidates, we were left with 12 undocumented alleles. Nine of the alleles were observed in more than a single individual's genotype, which increases our confidence in the inferences.

For further validation, we compared the undocumented alleles to three sources (Additional file 1: Table S5). In the first, an analysis of whole-genome short-read sequencing data in 286 individuals from Luo et al. [11], six out of the 12 undocumented alleles were observed. The second was a pmTRIG [25] data set, where we found four out of the 12. The third was a long-read whole-genome

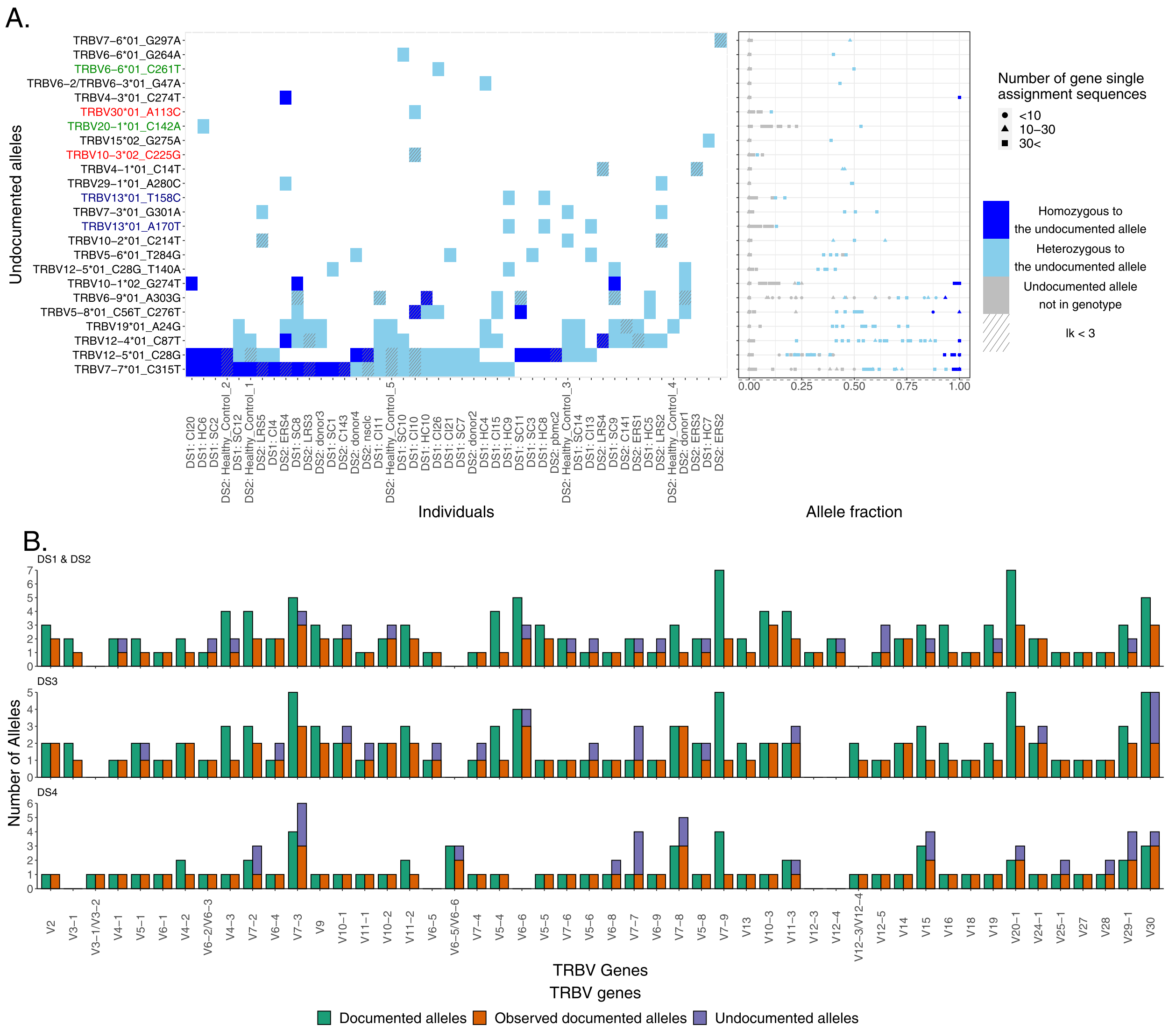

Fig. 2 Undocumented alleles. **A** Undocumented allele distribution in DS1 and DS2. The left panel shows undocumented alleles within the genotype. Each row represents an undocumented allele, and each column represents an individual. The *Y*-axis annotations in red, green, and blue correspond to alleles not following the expected multi-modal distribution, adjacent nucleotide stretches, or both, respectively. The tile colors correspond to an individual's genotype. The right panel shows the fraction of the undocumented allele assignments out of the gene assignments. The *X*-axis is the fraction and the *Y*-axis is the same as in **A**. Colors correspond to an individual's genotype of the allele. The shapes correspond to the number of gene assignments: a circle indicates that the number of gene assignments is less than 10, a triangle indicates that the number of gene assignments is between 10 and 30, and a square indicates that the number of gene assignments is more than 30. **B** Documented alleles versus the observed documented alleles and the undocumented alleles for each TRBV gene and in each data set, DS1 and DS2, DS3, and DS4, respectively. The *X*-axis corresponds to the TRBV genes and the *Y*-axis to the number of alleles. The color corresponds to the allele groups: dark red is the documented alleles, light red is the observed documented alleles, and blue is the observed undocumented alleles

sequencing data of 35 diploid donors from Ebert et al. [39] which confirmed six out of the 12. Altogether, eight of the 12 undocumented alleles were observed in at least one of these other data sets, and two were observed in all of the data sets. In addition to the 12 undocumented alleles, another five DS1 alleles matched known allele references in IMGT, but lacked 3′ nucleotides in IMGT (Additional file 1: Table S6). We replaced these IMGT alleles with their longer versions in our analyses, since it improves the genotype inference.

We also explored 5′ UTR genomic variation in DS1. Consensuses of the 5′ UTR were constructed as previously described in Mikocziova et al. [43, 44]. We compared the consensus sequences to the upstream regions that include leader 1 (L-PART1), leader 2 (L-PART2), and the upstream leader 1 sequence (Additional file 1: Table S7) and found 31 undocumented upstream sequences. The L-PART1 and L-PART2 of 10 of these sequences are absent from IMGT, and the L-PART1 of TRBV10-3*02 is also absent from IMGT (Additional file 1: Table S8). Four of the 11 absent L-PART1 sequences and six of the 10 absent L-PART2 sequences were also observed in long-read assemblies (Additional file 1: Table S8). Two sequence variants were observed in L-PART2. Both were associated with TRBV13*01 and were also observed in the long-read assemblies. In L-PART1, we found 15 sequence variants from 14 different alleles, and 13 of them were also observed in the long-read assemblies. In addition, we found four alternative splicing sequences from three different alleles. The first two are from clusters of TRBV23-1, an ORF gene that lacks a functional splice donor site [45]. As a result, the 5′UTR consensus sequence of TRBV23-1*01 contains an intron between L-PART1 and L-PART2. Both consensus sequences differ from the previously reported upstream sequence by the number of copies of the TTTTG motif (Additional file 1: Fig. S3). The other two alternative splicing variants were found upstream of the TRBV7-7 alleles. Here, both the documented *01 allele and the undocumented allele *01_C315T carry the alternative splicing sequences. Three more upstream inferred sequences that correspond to TRBV4-3*01, TRBV20-1*01, and TRBV20-1*02 are absent from IMGT. However, those three sequences match the reference of the TRB locus under GRCh38 [40]. The 3′ ends of the L-PART1 references of the three sequences in IMGT seem to originate from the intron, and the 5′ splice site of the introns of those three alleles were likely mis-identified in IMGT (Additional file 1: Fig. S4). We also found three variant consensus sequences associated with TRBV6-2*01/TRBV6-3*01 (Additional file 1: Fig. S3), all three were observed in the long-read assemblies for the TRBV6-2*01 annotation. One of them, TRBV6-2*01/TRBV6-3*01_2, was also observed in the undocumented allele TRBV6-3*01_G47A.

Next, we analyzed DS2, which includes data from 25 individuals. Fifteen undocumented alleles were inferred (Fig. 2A). Nine of them were also observed in DS1. Four others were present in the short-read whole-genome databases. One was in Luo et al. [11] and all four were in pmTRIG [25]. Two out of those four undocumented alleles were observed in the 35 diploid long-read assemblies from Ebert et al. [39]. One out of those two undocumented alleles was observed in all sources, and two of the 15 undocumented alleles were not observed in any of the sources, including DS1 (Additional file 1: Table S5).

Lastly, we investigated genomic variation in DS3 and DS4. Both data sets contain partial sequences, making it impossible to distinguish between alleles that differ in the regions outside those covered by the library primers. At first, we confirmed our inference approach by using artificial sequences trimmed from DS1 to the same sequence lengths as DS3 and DS4 (BIOMED-2 and Adaptive Biotechnologies, respectively; Additional file 1: Table S9). Of the 12 undocumented DS1 alleles, seven were in range for the BIOMED-2 primers and two for the Adaptive Biotechnologies primers. The artificial data sets enabled the inference of all the in-range undocumented alleles for both partial libraries. Having gained confidence in the approach, we then applied it to DS3 and DS4. From the 219 DS3 TRB genotypes [30], we inferred 29 potential undocumented allele patterns. However, 13 alleles failed the quality filtering steps (see the “Methods” section and Additional file 1: Table S10). Five of the remaining 16 undocumented alleles were independently identified in multiple genotypes, giving us added confidence in these inferences. Four alleles were supported by their identification by Luo et al. [11]. Six alleles were also supported by pmTRIG [25]. Eight alleles were not observed in any of the sources including DS1 and DS2 (Additional file 1: Table S10). A similar analysis of inferred genotypes from 786 individuals in the Adaptive Biotechnologies data set (DS4) identified a further 24 undocumented alleles (Additional file 1: Table S11). We discarded 10 alleles that failed the quality filtering steps (see the “Methods” section and Additional file 1: Table S11). Six out of the remaining 14 undocumented alleles were independently identified in multiple genotypes, giving us added confidence in these inferences. Two alleles were supported by the presence of alleles identified also by Luo et al. [11]. Five alleles were supported also by pmTRIG [25]. Eight alleles were not observed in any of the sources including DS1, DS2, and DS3 (Additional file 1: Table S11). To summarize this part, we compared the documented alleles for this locus with all alleles observed in our data sets and with the assemblies of genomic long reads (Fig. 2B). All in all we identified 39 undocumented TRBV alleles and 31 undocumented upstream sequences. Eight of those complete undocumented alleles and seven undocumented upstream sequences were also observed in the 35 long-read diploid assemblies of Ebert et al. [39].

Inference of double chromosome deletions in four TRBV genes

Deletion polymorphisms can be so common that an individual may carry the deletion in both chromosomes. We refer to these genes as double chromosome deletions, although it is important to note that they are deleted from the expressed repertoire and not necessarily from the genome itself (see the “Discussion” section). TRBV gene usage in DS1 shows such deletion polymorphisms in four genes (Fig. 3A). TRBV4-3 and TRBV3-2 are absent from the genotypes of eight individuals,

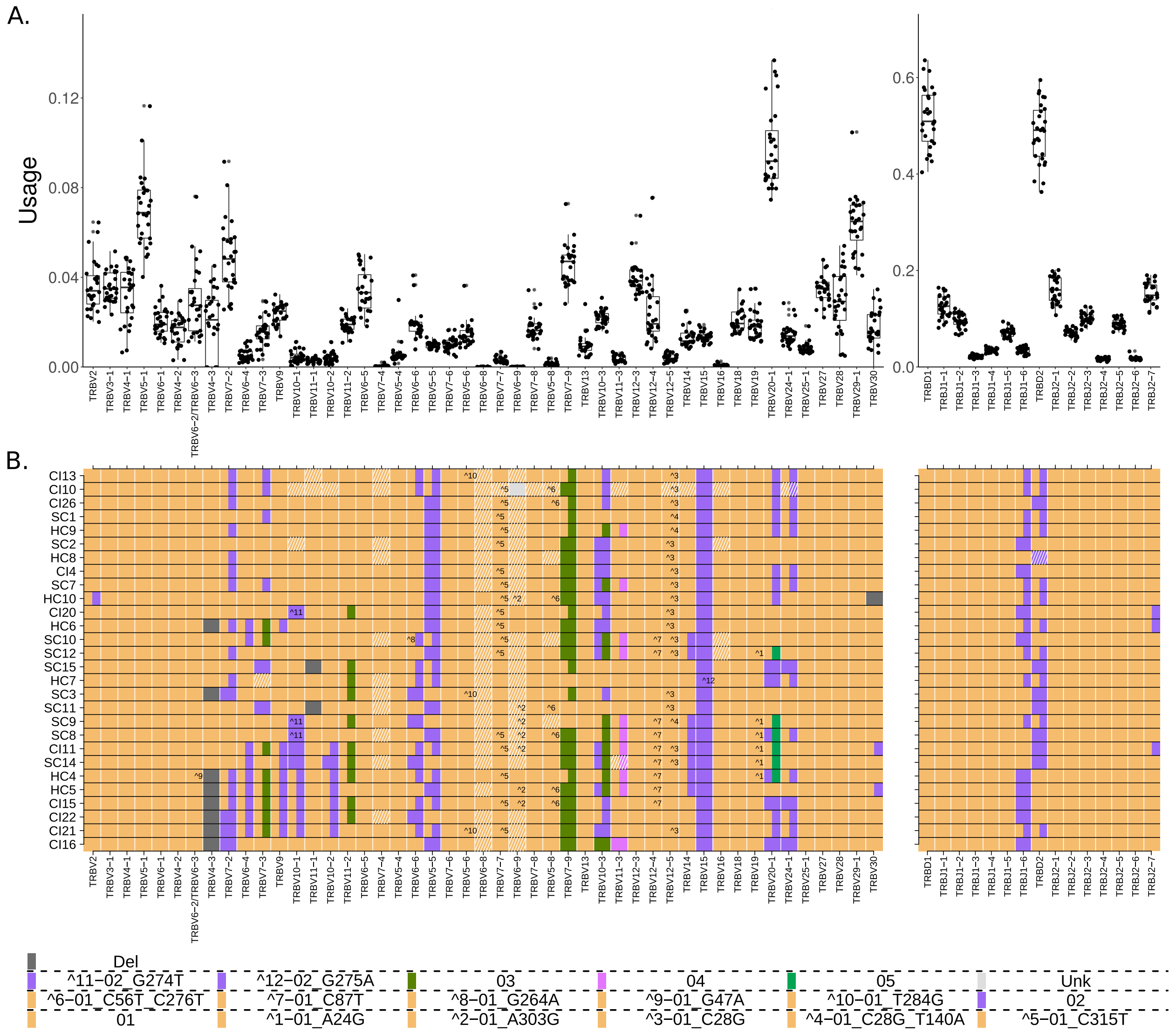

Fig. 3 Gene usage and genotypes of the DS1 data set. **A** TRB gene usage. Each dot is one individual's relative gene usage, which is the fraction of the number of sequences mapped to a given gene by the number of total sequences. The *X*-axis shows the TRB genes in the order in which they are found in the genome, except TRBV30 that is located after the TRBD and TRBJ section. The *Y*-axis shows the frequency of gene usage. **B** TRB genotypes. Each row shows an individual genotype, and columns correspond to the different genes in the same order as in **A**. Colors correspond to alleles as indicated in the bottom of the figure

while TRBV11-1 is absent from two individuals, and TRBV30 from one individual (Fig. 3B). In DS2, a similar data set to DS1 from the point of view of sequence length of the coding region, double chromosome deletions of TRBV4-3 were identified in eight individuals, TRBV3-2 in two individuals, and TRBV7-3 in one individual (Additional file 1: Fig. S5). Interestingly, the individual from DS1 with an inferred TRBV30 deletion (HC10) was shown to be homozygous or hemizygous for the undocumented allele TRBV30*03_T285C. On the assumption that this undocumented allele is found at relatively low frequency within the human population, homozygosity is unlikely. However, it is possible that the undocumented allele has escaped more widespread detection because of its low usage level. This low usage is a consequence of TRBV30*03_T285C being a pseudogene, because its coding region includes an in-frame stop codon. HC10 therefore has at least the functional equivalent of a double chromosome deletion.

The gene TRBV4-3 and the pseudogene TRBV3-2 were always inferred as being deleted together in DS1 individuals (Fig. 3B). TRBV4-3 and TRBV3-2 are close to each other, and a common ~21-kb deletion which includes TRBV4-3, TRBV3-2, and one of the genes

TRBV6-2/TRBV6-3 has been reported from genomic studies [10, 46–48]. The inference of a deletion of either TRBV6-2 or TRBV6-3 in AIRR-Seq data is made difficult, because the deletion might be hidden by the presence of the other identical TRBV6-2/TRBV6-3 gene. However, indirect evidence that the deletion polymorphisms seen in DS1 are associated with the previously reported ~21-kb deletion comes from the usage of TRBV6-2*01/TRBV6-3*01. In individuals who lack TRBV4-3 and TRBV3-2, usage of TRBV6-2*01/TRBV6-3*01 is significantly lower than in the individuals who express TRBV4-3 and TRBV3-2 (Additional file 1: Fig. S6). It is therefore likely that detection of TRBV6-2*01/TRBV6-3*01 in these individuals is entirely a consequence of sequences utilizing one of the genes. This line of reasoning also allowed us to conclude that an undocumented polymorphism seen in sample HC4 is most likely an allele of the undeleted gene, meaning it can be either TRBV6-3*01_G47A or TRBV6-2*01_G47A.

In DS2 individuals, deletion of V4-3 was not always accompanied by evidence of deletion of TRBV3-2. This is likely because DS2 was collected from different sources, and in some data sets, non-productive sequences had been filtered out. Evidence of the presence or absence of the TRBV3-2 pseudogene is therefore lacking. In other samples, analysis of TRBV3-2 usage is compromised by its low usage (Fig. 3A).

Strong asymmetry between the probabilities for mis-identification of TRBD2 alleles

There are only two TRBD genes, TRBD1 and TRBD2, and three reported TRBD sequences (TRBD1, TRBD2*01, and TRBD2*02). Both genes are short and highly similar. TRBD1 is 12bp long, and TRBD2 is 16bp long. Each sequence includes a short central motif flanked by G-rich ends. A single G/A SNP that is flanked by runs of Gs differentiates the two TRBD2 alleles (TRBD2*01: “GGGACTAGCGGG**G**GGG”, TRBD2*02:“GGGACTAG CGGG**A**GGG”). During V(D)J rearrangement, the ends of the TRBD segment are trimmed, and P-nucleotides and N-nucleotides are added between the joining TRBV, TRBD, and TRBJ genes [2]. Studies of N-nucleotide addition in BCR V(D)J genes show the process to be biased towards addition of Gs and addition of homopolymer tracts [49]. The unequivocal identification of germline-encoded nucleotides within the TCRβ V(D)J junctions is therefore problematic, and this is particularly true for the TRBD gene ends. To reduce errors, we limited the TRBD gene genotype analyses to sequences with a minimum inferred length of 9 bp. Some errors still remained, particularly as a result of TRBD2 allele assignment errors. Twenty-three out of the 28 individuals from DS1 were initially inferred to be heterozygous at the TRBD locus. However, this is unlikely according to the Hardy–Weinberg principle, which states that in equilibrium, in the absence of selection or other evolutionary pressures, the maximum frequency of heterozygous individuals in a population having two allelic variants of the gene is 0.5 [50].

To correct for these errors, we applied a process to determine the borders between TRBD2 genotype groups, as described in the “Methods” section (the “Determining the borders between TRBD2 genotype groups” section), resulting in a clear separation between the groups. This implies that TRBD2*01 is mis-identified as TRBD2*02 with a probability of 4%. The probability of mis-identifying TRBD2*01 as TRBD2*02 is estimated to be 12.7%. We can thus estimate the average usage of TRBD2*01 in heterozygous individuals corrected for these errors to be ~0.393. There is a strong asymmetry between the probabilities for mis-identification of TRBD2*01 as TRBD2*02 (GGGACTAGCGGG**G**GGG and GGGACTAGCGGG**A**GGG, respectively) or the opposite. Mis-identification is generally a result of several processes: (1) V(D)J recombination, where the ends of the TRBD segment are trimmed, and P-nucleotides and N-nucleotides are added between the joining TRBV, TRBD, and TRBJ genes ([49]); (2) PCR errors [51]; and (3) sequencing errors [38, 52]. All three contribute to this asymmetric mis-identification. It may also result from selection, from the amino acid differences in the sequences, or from unknown structural variants in the locus that could be associated with the different alleles.

Linkage disequilibrium between TRBJ1-6 and TRBD2

To explore J genotypes, we first checked for evidence of errors by exploring the fraction of all TRBJ1-6 assignments that are assigned to TRBJ1-6*01. This is the most likely error in TRBJ genotyping, as TRBJ1-6 is the only TRBJ gene with two known functional alleles. The distribution of frequencies shows a good partitioning between homozygous and heterozygous individuals (Additional file 1: Fig. S2D), indicating that the TRBJ1-6 alleles can be reliably inferred.

Of note, in heterozygous individuals, TRBJ1-6*02 is considerably more frequently used compared with TRBJ1-6*01 (Additional file 1: Fig. S2D). The average fraction of TRBJ1-6*01 out of all sequences assigned to TRBJ1-6 in heterozygous individuals is ~ 0.39, which is comparable with the average fraction of TRBD2*01 out of all sequences assigned to TRBD2 in TRBD2 heterozygous individuals after correcting for mis-assignments (see above). The similarity between the biased usage of TRBJ1-6 and TRBD2 alleles in heterozygous individuals led us to test the genetic dependency between these loci.

The distance between TRBJ1-6 to TRBD2 is relatively short (~6000bp), suggesting these loci could indeed be in linkage disequilibrium (LD). To test this hypothesis, we reviewed whole-genome sequencing (WGS) records from the 1000 Genomes Project to profile the region's variants based on TRBD2 haplotype (Additional file 1: Fig. S7). We observed SNPs with a high LD score (r-square) between the genes. Furthermore, the WGS haplotypes showed several other SNPs with high LD score scattered in the TRBD-TRBC2 genomic region, which strengthens the association between TRBD2 alleles and other markers in the locus. DS4 was unsuitable to test this hypothesis because DS4 sequences do not include the SNP that differentiates between TRBJ1-6*01 and TRBJ1-6*02. We therefore tested the LD hypothesis using DS3. Only genotypes for which we were confident of the TRBD2 genotype were taken into account. These genotypes are shown outside the gray areas of Additional file 1: Fig. S2B. Additional file 1: Fig. S8 shows that all of the homozygous TRBD2*01 individuals are also homozygous for TRBJ1-6*02. Also, 29 out of the 31 homozygous TRBJ1-6*01 individuals are homozygous for TRBD2*02.

TRBJ usage is strongly dependent on the TRBD2 genotype

Having confirmed that TRBJ1-6 and TRBD2 are in LD, we next investigated the influence of TRBD2 genotypes on TRBJ/TRBV gene usage in the repertoires. Since such an investigation requires accurate TRBD2 genotype inference, accurate annotations of TRBJ genes, and a large data set, DS4 was used.

We found that homozygous TRBD2*02 individuals tend to use TRBD2 1.29 times more than homozygous TRBD2*01 individuals (Fig. 4A, left panel). TRBD2 can undergo rearrangements only with TRBJ2 genes [53], so we expected that the usage of TRBJ2 genes should increase in homozygous TRBD2*02 individuals. Indeed, homozygous TRBD2*02 individuals use TRBJ2 genes significantly more than heterozygous and homozygous TRBD2*01 individuals (Fig. 4A, right panel), with 11 out of the 13 genes yielding *p* values lower than 0.001 (Mann–Whitney test, adjusted by Bonferroni correction). Furthermore, comparing the combined usage of all TRBJ2 genes to the combined usage of all TRBJ1 genes reveals a strong effect. For TRBD2*01 individuals, the mean usage of TRBJ1 was 0.473 compared to 0.366 for the TRBD2*02 individuals (Fig. 4B).

Next, we examined if and how TRBD2 haplotypes affect the relative usage of individual TRBJ1 and TRBJ2 genes. For this, we plotted the relative usage of individual TRBJ1 and TRBJ2 genes normalized independently for each gene family (Fig. 4C). Surprisingly, we found that TRBJ usage within each family is also affected by the TRBD2 genotype. Since TRBD2 can rearrange only with TRBJ2 genes, we stratified the above distributions into subsets that include only biologically possible rearrangements. In particular, Fig. 4D shows the conditional probability P(TRBJ2-N|TRBD2) for all the TRBJ2 genes. The biased usage of the TRBJ2 genes observed in Fig. 4C is still present, indicating that TRBD2 relative likelihood to recombine with TRBJ2 genes is strongly affected by the TRBD2 genotype. We then explored the TRBJ gene fraction of the sequences assigned to TRBD1 (P(TRBJ1/2-N|TRBD1), Fig. 4E) and observed that TRBD2 genotype is associated with the TRBD1 likelihoods to rearrange with individual TRBJ genes. We further investigated the effect of TRBD2 genotype on the likelihoods to rearrange with individual TRBJ2 genes only, and the effect was mostly eliminated in terms of magnitude P(TRBJ2-N|TRBD1), Fig. 4F).

The strong biases observed in Fig. 4A–E can result from amino acid alterations in the sequence, from non-coding regulatory variants, or from unknown structural variants in the locus that are associated with the different alleles. To discriminate between these options, we repeated the analysis for non-functional sequences that resulted from frame-shifts between the TRBV and TRBJ genes. Such non-functional sequences are commonly used to reflect the initial V(D)J usage prior to thymic selection [54–56]. Of note, these non-functional clones cannot overlap with functional clones, since in T cells clones are defined as sharing identical V(D)J sequence. In these non-functional sequences, the biases are pronounced in a similar fashion (Additional file 1: Fig. S9). Thus, we conclude that the differences between the TRBD and TRBJ rearrangements stratified by TRBD2 genotype are most likely due to structural differences or non-coding regulatory variants between the haplotypes rather than due to negative selection.

Haplotype inference reveals several association patterns

From the genotype analysis of DS1, we observed a potential pattern between homozygosity of TRBV7-2*02 and lack of usage of TRBV4-3 in four individuals. To inspect the link between TRBV7-2*02 and the lack of usage of TRBV4-3, we turned to haplotype inference. In an analogous way to the haplotyping methods for B cell receptor data [21, 22], a heterozygous TRBD or TRBJ gene was needed as an anchor for TRBV haplotype inference. A suitable candidate is TRBJ1-6, for which 11 individuals from DS1 (Fig. 3B) and eight individuals from DS2 (Additional file 1: Fig. S5) are heterozygous. In seven individuals who are heterozygous for TRBV7-2, the chromosome that carried the allele 02 of this gene had a clear deletion of TRBV4-3 (SAMPLE_ANCHOR-GENE_ALLELE: CI13_J1-6_01, SC12_J1-6_02, HC6_J1-6_01, HC9_J1-6_02, SC7_J1-6_02, LRS2_J1-6_01, and donor3_J1-6_02). The genotype and haplotype inferences both support the association between TRBV7-2 genotype and the usage of TRBV4-3. To further quantify this effect, we surveyed individual genotype and haplotype inferences in DS3. DS3 is a much larger data set, which

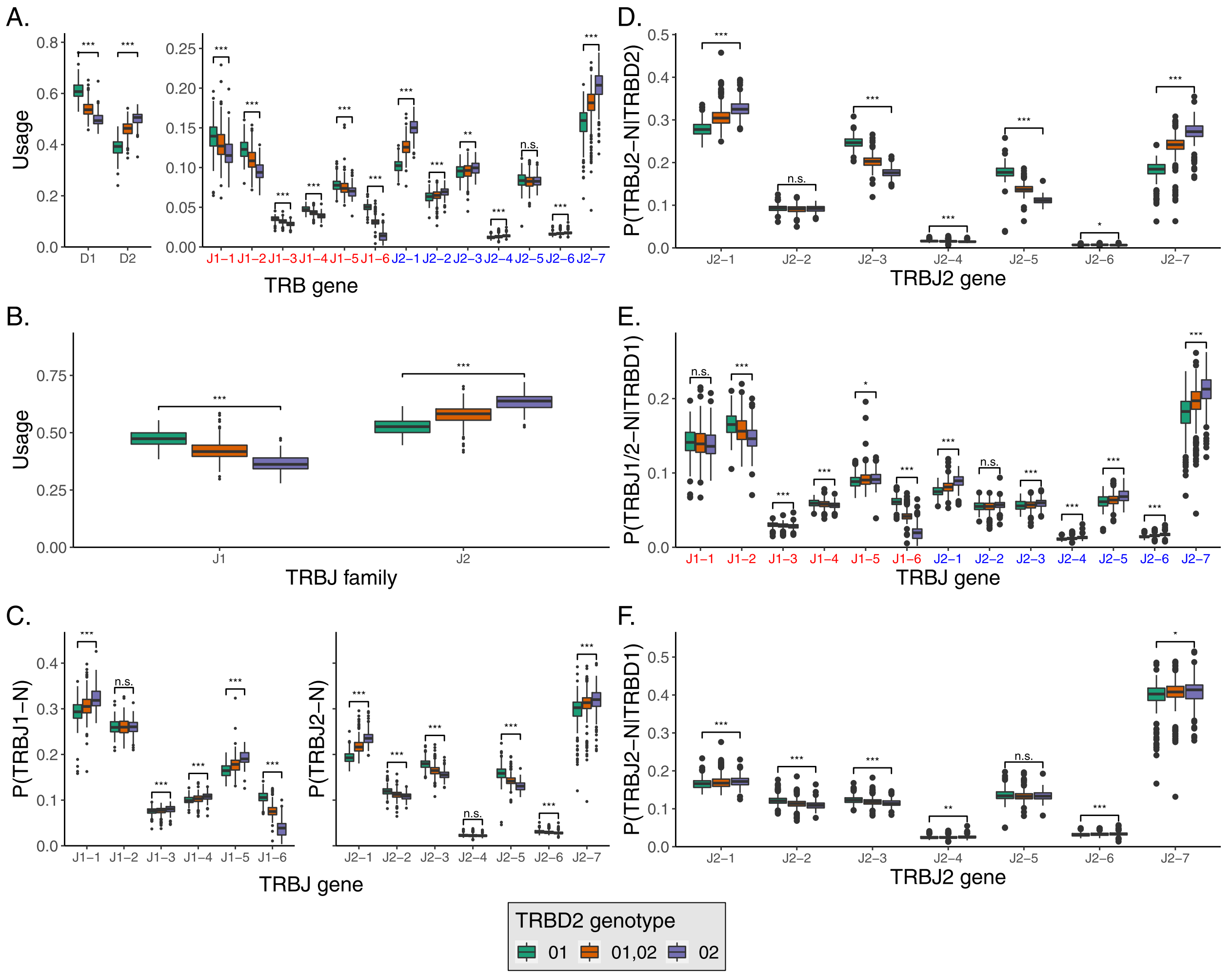

Fig. 4 TRBD and TRBJ usage correspond to TRBD2 genotype. **A** TRBD and TRBJ gene usage in DS4 individuals with different TRBD2 genotypes. TRBJ genes are shown along the *X*-axis in the order in which they appear in the genome. **B** The TRBJ family usage in DS4 individuals with different TRBD2 genotypes. TRBJ families are shown along the *X*-axis in the order in which they appear in the genome. **C** TRBJ gene usage normalized to the TRBJ family usage, in DS4 individuals with different TRBD2 genotypes. TRBJ genes are shown along the *X*-axis in the order in which they appear in the genome. **D** The relative fraction of TRBJ2 genes out of the sequences that were assigned to TRBD2 and were longer than 7nt. **E** The fraction of TRBJ2 genes out of the sequences that were assigned to TRBD1 and were longer than 7nt. TRBJ genes are shown along the *X*-axis in the order in which they appear in the genome. **F** The fraction of TRBJ genes out of the sequences that were assigned to TRBD1 and were longer than 7nt. TRBJ genes are shown along the *X*-axis in the order in which they appear in the genome. The box colors correspond to the TRBD2 genotype. Statistical significance was determined using a Mann–Whitney test and adjusted by Bonferroni correction (*n.s.* not significant, *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$)

covers one of the unique SNPs of TRBV7-2*02, allowing the differentiation of TRBV7-2*02 from the rest of the known TRBV7-2 alleles (Additional file 1: Table S2). In eight out of nine individuals with a high genotype inference likelihood, the pattern between homozygosity of TRBV7-2*bp02 (see the section "Allele pattern collapsing") and a deletion inference of TRBV4-3 was apparent (Additional file 1: Fig. S10 and S11). Another gene with a link to TRBV7-2 is TRBV6-2/TRBV6-3. Its usage was also highly affected by the genotype of TRBV7-2. The mean usage of TRBV6-2/TRBV6-3 in TRBV7-2*bp02 homozygous individuals was less than half of the mean usage in TRBV7-2*bp01 homozygous individuals. Since we cannot distinguish between TRBV6-2 and TRBV6-3, this observation supports the hypothesis that both TRBV4-3 and one of TRBV6-2/TRBV6-3, are not present in haplotypes that carry TRBV7-2*bp02. We used long-read assembly haplotypes to investigate the connection between TRBV7-2 and the deletions. All 40 haplotypes that carried TRBV7-2*bp02 had a deletion stretch at the genomic location of TRBV4-3 and either TRBV6-2 or TRBV6-3. An additional 27 haplotypes carried TRBV7-2*bp01, of which eight had the same deletion stretch as with TRBV7-2*bp02.

In addition, the following association patterns between specific alleles and single chromosome deletions were revealed by haplotype inference: (1) TRBV24-1*02 was observed in eight samples on the chromosome carrying TRBJ1-6*02 and none in the other chromosome. (2) TRBV28*Del was observed in 4 samples on the chromosome carrying TRBJ1-6*02 and none in the other chromosome. (3) TRBV20-1*05 was observed in five samples on the chromosome carrying TRBJ1-6*01 and none in the other chromosome. (4) In eight individuals when TRBV24-1*02 was present on chromosome TRBJ1-6*02, TRBV20-1*02 was also observed. Of note is a haplotype block between the genes TRBV6-4 to TRBV10-1 that was observed in DS1:CI21 on the TRBJ1-6*01 chromosome and in DS2:donor4 on the TRBJ1-6*02 chromosome (Fig. 5).

TRBD2 was also considered a potential anchor gene for haplotype inference. To examine the reliability of this inference, we used the same individuals who were heterozygous for TRBJ1-6. Ten out of the 11 individuals from DS1 were heterozygous for TRBD2 and were used for the haplotype inference (Additional file 1: Fig. S12). Although the number of recombinations of TRBD2 with the TRBV genes is much larger than TRBJ1-6 and could potentially supply a better inference, comparison of the results from both anchor genes shows a different picture. The haplotypes inferred with TRBD2 commonly show occurrences of more than one allele per gene on a single chromosome. This is most likely due to ambiguous assignment of the very short and similar TRBD2 alleles. Hence, although haplotype inference with TRBD2 is feasible, it is likely to be less accurate.

Discussion

In this study, we have explored how beta chain TCR V(D)J sequences of different lengths, generated using differing

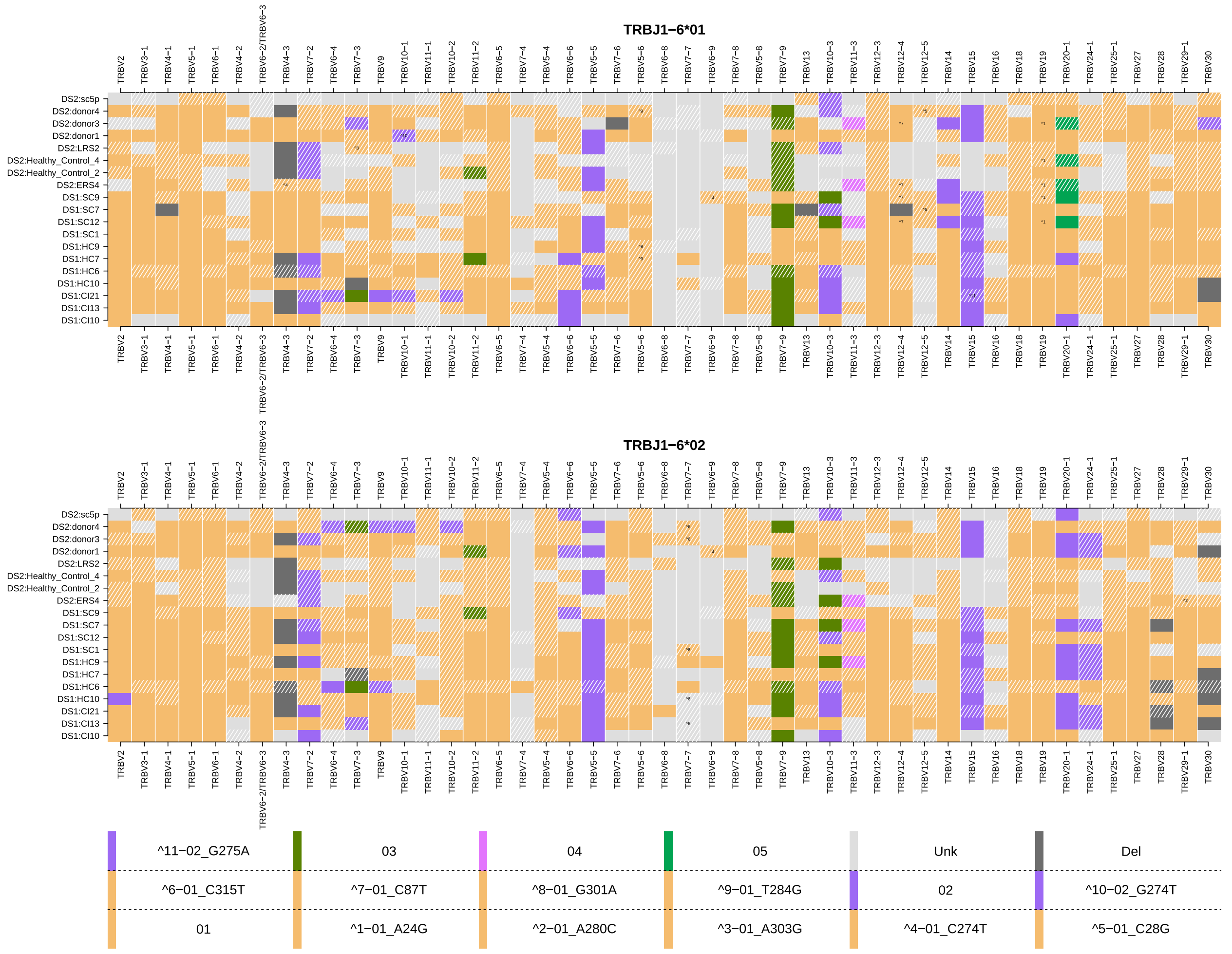

Fig. 5 TRBV haplotypes for 19 individuals from DS1 and DS2. The upper and lower panels show the TRBV haplotypes anchored with TRBJ1-6*01 and TRBJ1-6*02, respectively. Each row is an individual's haplotype, and each column is a V gene call. The colors correspond to the V alleles and the tile annotations correspond to the undocumented allelic variants

technologies, can contribute to our knowledge of TRB genes, their allelic variants, and their influence on the expressed repertoire. Previous studies of BCR genotypes and haplotypes have led to the identification of dozens of new allelic variants of heavy and light chain variable region genes. This approach has not been extended to investigations of the TCR genes. Zhang et al. [57] investigated the reliability of V gene identification for different lengths of sequences of IGH, IGK, IGL, TRA, and TRB with respect to somatic hypermutations and sequencing errors. The comparison there is between full-length and lengths of 100, 150, or 200 nucleotides, whereas we used an analogous approach to compare the lengths of the common sequencing protocols. The focus of most TCR gene studies remains firmly fixed on the CDR3 regions of the genes [58]. Despite their interaction with the major histocompatibility complex (MHC), and their documented influences on TCR/MHC/peptide interactions [59, 60], the CDR1 and CDR2 gene sequences and their translated products are still generally ignored in TCR repertoire studies. Only the 3′ ends of the TCR variable gene sequences are included in the amplicons generated by commercial providers of TCR sequencing such as BGI, iRepertoire, and Adaptive Biotechnologies. Their sequencing setup allows the unambiguous identification of most variable region genes that may partially encode the CDR3 sequences, but they are rarely able to identify TRBV genes at the allele level. This study sheds light on understudied regions of the TCR, to enable accurate identification of new alleles, genotypes, and haplotypes.

Despite the intimate partnership of MHC proteins and TCRs in the recognition of antigenic peptides, it is only the genes of the MHC that are widely recognized as disease susceptibility genes [61]. The germline genes that rearrange to produce TCRs are rarely accorded much importance. The multiple sets of TCR alpha, beta, gamma, and delta V, D, and J genes each include many highly similar genes. This may have encouraged the view that the astonishing processes of V(D)J recombination should generate much the same kind of repertoire, no matter which germline genes are available to an individual.

Germline TCR genes may also represent a blind-spot to the immunological community, because until recently they were so difficult to document in an individual, let alone in a population. High-throughput sequencing studies of TCR repertoires now enable easy documentation of germline genes. In this study, we adapt tools and techniques that were developed for analysis of BCR genotypes [18, 19, 21, 23, 62, 63] and haplotypes [17, 22] for the analysis of TCR data sets. Other tools such as LymAnalyzer have been specifically developed for the analysis of TCR data [64], but they do not produce genotypes and haplotypes.

Our analysis demonstrates that many undocumented V genes remain to be discovered and that even partial AIRR sequences can be analyzed for the detection of undocumented polymorphisms. It is clear, however, that much additional information is to be gained by the study of full-length V(D)J genes. Analysis of genotypes and haplotypes from full-length sequence data sets of 53 individuals led to the identification of 18 TRBV alleles that are not documented in the IMGT Reference Directory. In contrast, only 16 polymorphisms were identified in truncated sequences generated from 219 individuals using BIOMED-2 primers, and just 14 new allele patterns were seen in the 786 Adaptive Biotechnologies data sets. This latter result reflects both the very short lengths of the Adaptive Biotechnologies sequences and the general lack of variability in the 3′ ends of the TRBV genes. To try and speculate whether the TRB polymorphisms are more important for MHC binding or antigen binding, we compared the distributions of known SNPs in TRBV and IGHV genes (Additional file 1: Fig. S13 and Fig. S14), expecting to find more amino acid replacements in the regions of CDR1/2. Schwartz et al. [65] have shown that germline diversity at a given position is a good indicator for the potential to survive after a somatic mutation at that position. However, there are far fewer known SNPs in TRBV than in IGHV, which makes it impossible to address this interesting question. We also could not address the question of polymorphisms in TRA, as there are very few full-length TRA samples available.

It is interesting to contrast this level of gene discovery with genetic variation amongst the BCR-encoding genes: to date, 120 functional alleles are documented in IMGT for 48 TRBV genes (avg. 2.5 alleles per gene). This compares to 286 functional alleles reported for the 55 IGHV genes (avg. 5.2 alleles per gene). This study adds a total of 39 alleles and SNPs to the record of TRBV genes (∼ 33% of the known alleles), and it seems likely that many undocumented TRBV alleles remain to be discovered.

It appears that there is less structural variation in the TRBV locus than is seen in the IGHV locus. A well-documented 21-kb deletion polymorphism in the TRBV locus, involving the TRBV4-3, TRBV3-2, and TRBV6-2 genes, was frequently noted here. Other deletion polymorphisms, each involving single genes — TRBV11-1, TRBV7-3, TRBV30, TRBV2, TRBV4-2, TRBV6-5, TRBV5-5, TRBV13, and TRBV28 — were seen at relatively low frequencies. No deletion polymorphisms involving the TRBD or TRBJ loci were detected. In contrast, numerous relatively common deletions are now known in the IGH loci, including one involving 13 consecutive, functional IGHV genes, and one involving six consecutive IGHD genes [21, 23]. The seven functional IGHV genes that can be found between IGHV4-28 and IGHV4-34 are rarely all present, or all absent, with at least six recognized structural variant haplotypes [16, 23].

It should be emphasized that the deletions reported here reflect an absence of rearrangements in the expressed TCR repertoire. It is possible that the deletions are "functional" rather than structural, perhaps as a result of variants in recombination signal sequences (RSS) or other regulatory elements. Certainly in a few individuals, a handful of rearrangements of the genes in question were seen. For example, TRBV4-3 was usually present in about 2% of all rearrangements, but in eight individuals, it was seen at frequencies less than 0.05%.

No previously unknown gene duplications were inferred in this study. This is in sharp contrast to observations from studies of the IGHV locus. Early restriction fragment length polymorphism (RFLP)-based analyses pointed to duplications of sequences that came to be known as IGHV3-23 and IGHV1-69 [66–70]. More recently, the duplication of these two genes and of three other functional genes was confirmed, first by analysis of AIRR-Seq data [71] and then from genomic assembly data, including the second complete assembly of the IGHV locus [14].

It was not possible in this study to explore possible RSS variants, as RSS are lost from the genome during V(D)J recombination. On the other hand, 5′ RACE data allows for the exploration of variation in the 5′ UTR, as has recently been reported from BCR repertoire studies [43, 44]. Thirty-one variants of the 5′ UTR sequences of TRBV genes were identified in this study — a similar level of variability to that seen in IGHV studies. The functional implications of this kind of variation is yet to be determined.

The documentation of allelic variation and structural variation in the TCR gene loci will be important, as there are clear consequences of such variation on the expressed TCR repertoire, and if such variation can be shown to have consequences for the disease susceptibility of different individuals. It might be assumed that thymic selection would so powerfully shape individual TCR repertoires that any consequences upon the TCR repertoire of individual genotypes and haplotypes would be obscured. This study shows, however, that the usage of particular genes in the expressed repertoire appears to be very similar between individuals. The "shape" of the TCR repertoire may therefore be as predictable as has been found for the BCR repertoire [3, 4], reflecting both the carriage of individual genes and the LD that is found within the loci. Conspicuous LD identified in this study includes that of the TRBV4-3/TRBV3-2/(TRBV6-2 or TRBV6-3) deletion polymorphism and carriage of the TRBV7-2*02 allele, as well as linkage between the TRBD2 and TRBJ1-6 loci. These different haplotypes, in turn, are associated with significant differences in the usage of neighboring genes.

The power of AIRR-Seq analysis is well demonstrated by the TRBD gene analysis in this study. Although most TCR AIRR-Seq studies have been CDR3-focused, the TRBD genes that provide recurring central motifs to the CDR3 have usually been ignored. This study demonstrates that meaningful analysis of TRBD genes is possible and that even the slight sequence variation between the TRBD genes have consequences for the expressed repertoire. Within large V(D)J data sets, TRBD genes can be identified with confidence, and even the presence of different TRBD2 alleles in an individual's genotype can shape the expressed repertoire in predictable ways. Yet, all of these findings were procured mainly from individuals from developed countries or undocumented geographic origin. Sequencing individuals from developing and least developed countries could very well add or alter some of these results, as argued by Peng et al. [72]. Nonetheless, as our knowledge of these and other genes of the TCR loci grows, the way will be open to identify "departures from the repertoire norm" that may have biological and perhaps even clinical implications.

The frequency distribution analysis shown in Fig. 2A focuses on bi- or tri-modal distributions. In cases that involve duplicated genes that share the same allele, such as TRBV6-2 and TRBV6-3, having more than two alleles per gene may shift the location of the second/third/fourth mode of the distribution, but it will still not be uni-modal. As such, as long as we do not have extreme cases in which the relative frequency of the candidate allele is comparable to the mis-identification rate of the allele, the method is applicable. In other cases, such as the ones present in the IGH/IGK loci, these assumptions should be revisited [73].

To date, there have been very few associations of TCR genes with disease. One clear example is the genetic predisposition to carbamazepine-induced Stevens–Johnson syndrome (SJS), a severe cutaneous hypersensitivity with high mortality [74]. SJS and other cutaneous hypersensitivity reactions have been linked to HLA types, but these associations have all had a low positive predictive value [75], leading others to explore a possible role for TCR genes. It has now been shown that SJS is associated with the usage by cytotoxic T cells of a public TCR clonotype encoded by the TRBV12-4 and TRBJ2-2 genes [76]. Full-length sequences were not reported in the study of Pan and colleagues [76], and so a possible role for specific allelic variants of these genes could not be explored. Interestingly, in the present study, a previously undocumented polymorphism of TRBV12-4 was identified.

The lack of disease associations with TCR genes is likely to be a reflection of our ignorance of individual genetic variation within the TCR loci. Only after thorough exploration of the population genetics of the TCR genes, and of individual variation in the expressed TCR repertoire, will it be possible to determine whether or not these genes have a role in disease susceptibility. Genomic sequencing of TCR genes will contribute to this [77], but the present study demonstrates that it will also be possible to do this efficiently through the analysis of AIRR-Seq data. For this reason, the amplification of full-length TCR V(D)J sequences, or genomic long-read paired sequencing of the locus, must be strongly encouraged in AIRR-Seq studies. The results presented in this paper pave the way towards establishing functional links between the TRB germline repertoire and the TCR immune response. They expand our knowledge of genomic variation in the TRB locus and lay the ground for more accurate basic and clinical studies.

Conclusions

To summarize our findings, we identified 39 undocumented TRBV alleles and 31 undocumented upstream sequences, and inferred double chromosome deletions in TRBV4-3, TRBV3-2, TRBV11-1, and TRBV30. For TRBD and TRBJ, we determined the error rates in identification between TRBD2*01 and TRBD2*02 and corrected them, discovered LD between TRBJ1-6 and TRBD2, and found a strong bias in the usage of TRBJ genes depending on which TRBD2 allele is used. For example, for TRBD2*01 individuals, the mean usage of TRBJ1 was 0.473 compared to 0.366 for the TRBD2*02 individuals. Overall, our study sheds light on the genomic loci encoding TRB, to enable identification of new alleles, genotypes, and haplotypes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-01008-4.

Additional file 1: Supplementary materials: TRB germline variability is revealed by inference from repertoire data Supplementary figures 1–14: **Figure S1** Unique sequence distribution in the four data sets. **Figure S2** TRBD2 and TRBJ1-6 genotypes and gene usage frequencies. **Figure S3** 5′ UTR nucleotide sequences of TRBV genes. **Figure S4** 5′ UTR nucleotide sequences of TRBV genes. **Figure S5** DS2 TRB genotype heatmap. **Figure S6** TRBV6-2/TRBV6-3 usage correlates to the existence of TRBV4-3and TRBV3-2 in DS1. **Figure S7** Linkage disequilibrium in TRBD2 haplotype. **Figure S8** Genomic correlation between TRBD2 and TRBJ1-6 in DS3. **Figure S9** TRBD and TRBJ usage out of the non-functional sequences corresponds to TRBD2 genotype in DS4. **Figure S10** Heatmap of TRBV4-3 genotype correlates to the TRBV7-2 genotype in DS3. **Figure S11** TRBV gene usage according to the TRBV7-2 genotype in DS3. **Figure S12** TRBV haplotypes for 10 individuals from DS1. **Figure S13** TRBV SNPs distribution. **Figure S14** IGHV SNPs distribution. Supplementary tables 1–11: **Table S1** DS2 data set sources citation. **Table S2** BIOMED-2 allele patterns. **Table S3** Adaptive allele patterns. **Table S4** The distribution parameters of the TRBD2*01 fraction accordingto the TRBD2 genotype group in DS4. **Table S5** Previously unknown alleles comparison. **Table S6** Incomplete allele extensions table. **Table S7** 5′ UTR variants. **Table S8** 5′ UTR undocumented sequences. **Table S9** Undocumented allele verification in artificial partial libraries. **Table S10** Previously unknown alleles comparison. **Table S11** Previously unknown alleles comparison.

Acknowledgements

We would like to thank Mats Ohlin, Martin Corcoran, and James Heather for stimulating discussions.

Authors' contributions

GY conceived the idea, analyzed the data, and supervised the work. AO and AP developed the methods and analyzed and interpreted the data. OLR and CTW analyzed the genomic data. PP, AO, AP, AMC, WL, and GY wrote the paper. All authors read and approved the final manuscript.

Funding

This study was partially supported by grants from the ISF (832/16 to GY) and the European Union's Horizon 2020 research and innovation program (825821). The contents of this document are the sole responsibility of the iReceptor Plus Consortium and can under no circumstances be regarded as reflecting the position of the European Union. OLR and CTW were supported in part by funding from the National Institute of Allergy and Infectious Diseases (R24AI138963 to CTW).

Availability of data and materials

Alleles inferred in this study are listed in Additional file 1 (Tables S5-S6 and Tables S10-S11). The code for the analyses and the results are available in the Github repository (www.github.com/omaviv/TCR_genotype) [78]. For repertoire analyses of the four data sets, and the sequences of the undocumented alleles inferred, please refer to www.github.com/omaviv/TCR_genotype/tree/master/figures/data. For IMGT TRBV references used, please refer to www.github.com/omaviv/TCR_genotype/tree/master/pipeline/fasta_references. The genotypes and haplotypes are also accessible through VDJbase [79] (www.vdjbase.org); in case the reader needs help accessing the files, please refer to the help page on (https://vdjbase.org/user-guide).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Engineering, Bar Ilan University, 5290002 Ramat Gan, Israel. [2]Bar Ilan institute of Nanotechnology and Advanced Materials, Bar Ilan University, 5290002 Ramat Gan, Israel. [3]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, USA. [4]Institute of Structural and Molecular Biology, Birkbeck College, University of London, London, UK. [5]School of Biotechnology and Biomedical Sciences, University of New South Wales, Sydney, Australia.

Received: 17 June 2021 Accepted: 8 December 2021
Published online: 07 January 2022

References

1. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014;32(2):158–68.
2. Murphy K, Weaver C. Janeway's immunobiology. New York: Garland science; 2017.
3. Glanville J, Kuo TC, von Büdingen H-C, Guey L, Berka J, Sundar PD, Huerta G, Mehta GR, Oksenberg JR, Hauser SL, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci. 2011;108(50):20066–20071.
4. Rubelt F, Bolen CR, McGuire HM, Vander Heiden JA, Gadala-Maria D, Levin M, Euskirchen GM, Mamedov MR, Swan GE, Dekker CL, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. Nat Commun. 2016;7(1):1–12.
5. Collins AM, Yaari G, Shepherd AJ, Lees W, Watson CT. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? Curr Opin Syst Biol. 2020;24:100–8.
6. Hedrick SM, Nielsen EA, Kavaler J, Cohen DI, Davis MM. Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins. Nature. 1984;308(5955):153–158.

7. Hayday AC, Diamond DJ, Tanigawa G, Heilig JS, Folsom V, Saito H, Tonegawa S. Unusual organization and diversity of T-cell receptor a-chain genes. Nature. 1985;316(6031):828–832.
8. Yoshikai Y, Clark SP, Taylor S, Sohn U, Wilson BI, Minden MD, Mak TW. Organization and sequences of the variable, joining and constant region genes of the human T-cell receptor α-chain. Nature. 1985;316(6031): 837–840.
9. Toyonaga B, Yoshikai Y, Vadasz V, Chin B, Mak TW. Organization and sequences of the diversity, joining, and constant region genes of the human T-cell receptor beta chain. Proc Natl Acad Sci. 1985;82(24):8624–8628.
10. Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Barbié V, Chaume D. IMGT, the international immunogenetics database. Nucleic Acids Res. 1999;27(1):209–212. 10.1093/nar/27.1.209. /oup/backfile/content_public/journal/nar/27/1/10.1093/nar/27.1.209/2/27-1-209.pdf.
11. 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):. https://doi.org/10.26508/lsa.201800221. http://www.life-science-alliance.org/content/2/2/e201800221.full.pdf.
12. Mackelprang R, Livingston RJ, Eberle MA, Carlson CS, Yi Q, Akey JM, Nickerson DA. Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. Hum Genet. 2006;119(3):255–266.
13. Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA. Sequence variation and linkage disequilibrium in the human T-cell receptor β (TCRB) locus. Am J Hum Genet. 2001;69(2):381–395.
14. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, Wilson RK, Holt RA, Eichler EE, Breden F. 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(4):530–546.
15. Ford M, Haghshenas E, Watson CT, Sahinalp SC. Genotyping and copy number analysis of immunoglobin heavy chain variable genes using long reads. Iscience. 2020;23(3):100883.
16. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, Deikus G, Auckland K, Eichler EE, Marasco WA, Sebra R, Sharp AJ, Smith ML, Bashir A, Watson CT. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. Front Immunol. 2020;11:2136. https://doi.org/10.3389/fimmu.2020.02136.
17. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, Martin M, Hedestam GBK. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun. 2016;7(1):13642. https://doi.org/10.1038/ncomms13642.
18. 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. 2015;112(8):862–870.
19. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, O'Connor KC, Yaari G, Kleinstein SH. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. Front Immunol. 2019;10:129. https://doi.org/10.3389/fimmu.2019.00129.
20. Ralph DK, Matsen IV FA. Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. PLoS Comput Biol. 2019;15(7): 1007133.
21. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, Fire AZ, Tanaka MM, Gaëta BA, Collins AM. 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):1333–1340. https://doi.org/10.4049/jimmunol.1102097. https://www.jimmunol.org/content/188/3/1333.full.pdf.
22. Peres A, Gidoni M, Polak P, Yaari G. RAbHIT: R antibody haplotype inference tool. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz481. http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz481/28863273/btz481.pdf.
23. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, Sarna VK, Lundin KE, Clouser C, Vigneault F, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat Commun. 2019;10(1):1–14.
24. Yu Y, Ceredig R, Seoighe C. A database of human immune receptor alleles recovered from population sequencing data. J Immunol. 2017;198(5):2202–2210.
25. Khatri I, Berkowska MA, van den Akker EB, Teodosio C, Reinders MJT, van Dongen JJM. Population matched (PM) germline allelic variants of immunoglobulin (IG) loci: new pmIG database to better understand IG repertoire and selection processes in disease and vaccination. bioRxiv. 2020. https://doi.org/10.1101/2020.04.09.033530. https://www.biorxiv.org/content/early/2020/04/10/2020.04.09.033530.full.pdf.
26. Watson CT, Matsen 4th FA, Jackson KJ, Bashir A, Smith ML, Glanville J, Breden F, Kleinstein SH, Collins AM, Busse CE. Comment on "a database of human immune receptor alleles recovered from population sequencing data". J Immunol (Baltim: 1950). 2017;198(9):3371–3373.
27. van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, García-Sanz R, van Krieken JHJM, Droese J, González D, Bastard C, White HE, Spaargaren M, González M, Parreira A, Smith JL, Morgan GJ, Kneba M, Macintyre EA. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia. 2003;17(12):2257–2317. https://doi.org/10.1038/sj.leu.2403202.
28. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009;114(19): 4099–4107. https://doi.org/10.1182/blood-2009-04-217604. PMC2774550[pmcid].
29. Eliyahu S, Sharabi O, Elmedvi S, Timor R, Davidovich A, Vigneault F, Clouser C, Hope R, Nimer A, Braun M, Weiss YY, Polak P, Yaari G, Gal-Tanamy M. Antibody repertoire analysis of hepatitis C virus infections identifies immune signatures associated with spontaneous clearance. Front Immunol. 2018;9:3004. https://doi.org/10.3389/fimmu.2018.03004.
30. Simnica D, Akyüz N, Schliffke S, Mohme M, v.Wenserski L, Mährle T, Fanchi LF, Lamszus K, Binder M. T cell receptor next-generation sequencing reveals cancer-associated repertoire metrics and reconstitution after chemotherapy in patients with hematological and solid tumors. Oncoimmunology. 2019;8(11):1644110. https://doi.org/10.1080/2162402X.2019.1644110. PMID: 31646093. https://doi.org/10.1080/2162402X.2019.1644110.
31. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, Rieder M, Robins HS. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet. 2017;49(5):659–665. https://doi.org/10.1038/ng.3822.
32. 10x Genomics. 10X datasets. https://support.10xgenomics.com/single-cell-vdj/datasets. Accessed 15 Dec 2020.
33. Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, Liu X, Xie L, Li J, Ye J, Dong L, Cui X, Miao Y, Wang D, Dong J, Xiao C, Chen W, Wang H. Immune cell profiling of COVID-19 patients in the recovery stageby single-cell sequencing. Cell Discov. 2020;6(1):31. https://doi.org/10.1038/s41421-020-0168-9.
34. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, Liu L, Amit I, Zhang S, Zhang Z. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med. 2020;26(6):842–844. https://doi.org/10.1038/s41591-020-0901-9.
35. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, Knoetze N, Breden FMW, Christley S, Scott JK, Cowell LG, Breden F. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev. 2018;284(1): 24–41.
36. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, Vigneault F, Kleinstein SH. presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics. 2014;30(13):1930–1932. 10.1093/bioinformatics/btu138. /oup/backfile/content_public/journal/bioinformatics/30/13/10.1093_bioinformatics_btu138/2/btu138.pdf.
37. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. Bioinformatics. 2015;31(20): 3356–3358. 10.1093/bioinformatics/btv359. /oup/backfile/content_public/journal/bioinformatics/31/20/10.1093/bioinformatics/btv359/3/btv359.pdf.
38. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 2021;3(1):019.

39. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372(6537):eabf7117.
40. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood J, Clarke L, Koren S, Boitano M, Li H, Chin C-S, Phillippy AM, Durbin R, Wilson RK, Flicek P, Church DM. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. bioRxiv. 2016. https://doi.org/10.1101/072116. https://www.biorxiv.org/content/early/2016/08/30/072116.full.pdf.
41. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13(1):1–18.
42. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–26.
43. Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, Sollid LM. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. Nucleic Acids Res. 2020;48(10):5499–5510. https://doi.org/110.1093/nar/gkaa310. https://academic.oup.com/nar/article-pdf/48/10/5499/33326546/gkaa310.pdf.
44. Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM. Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. Iscience. 2021;24(10):103192.
45. Manfras BJ, Terjung D, Boehm BO. Non-productive human TCRβ chain genes represent V-D-J diversity before selection upon function: insight into biased usage of TCRBD and TCRBJ genes and diversity of CDR3 region length. Hum Immunol. 1999;60(11):1090–1100. https://doi.org/110.1016/S0198-8859(99)00099-3.
46. Zhao TM, Whitaker SE, Robinson MA. A genetically determined insertion/deletion related polymorphism in human T cell receptor beta chain (TCRB) includes functional variable gene segments,. J Exp Med. 1994;180(4):1405–1414. https://doi.org/10.1084/jem.180.4.1405. https://rupress.org/jem/article-pdf/180/4/1405/1105782/1405.pdf.
47. Rowen L, Koop BF, Hood L. The complete 685-kilobase dna sequence of the humanβ t cell receptor locus. Science. 1996;272(5269):1755–1762. https://doi.org/10.1126/science.272.5269.1755. https://science.sciencemag.org/content/272/5269/1755.full.pdf.
48. Brennan RM, Petersen J, Neller MA, Miles JJ, Burrows JM, Smith C, McCluskey J, Khanna R, Rossjohn J, Burrows SR. The impact of a large and frequent deletion in the human TCRβ locus on antiviral immunity. J Immunol. 2012;188(6):2742–2748. https://doi.org/10.4049/jimmunol.1102675. https://www.jimmunol.org/content/188/6/2742.full.pdf.
49. Jackson KJ, Gaëta BA, Collins AM. Identifying highly mutated IGHD genes in the junctions of rearranged human immunoglobulin heavy chain genes. J Immunol Methods. 2007;324(1-2):26–37.
50. Andrews C. The Hardy-Weinberg principle. Nat Educ Knowl. 2010;3(10):65.
51. Clarke L, Rebelo C, Goncalves J, Boavida M, Jordan P. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol Pathol. 2001;54(5):351.
52. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep. 2018;8(1):1–14.
53. Wallace ME, Bryden M, Cose SC, Coles RM, Schumacher TN, Brooks A, Carbone FR. Junctional biases in the naive TCR repertoire control the CTL response to an immunodominant determinant of HSV-1. Immunity. 2000;12(5):547–556. https://doi.org/110.1016/s1074-7613(00)80206-x.
54. Cui A, Di Niro R, Vander Heiden JA, Briggs AW, Adams K, Gilbert T, O'Connor KC, Vigneault F, Shlomchik MJ, Kleinstein SH. A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. J Immunol. 2016;197(9):3566–3574.
55. Yaari G, Vander Heiden J, Uduman M, Gadala-Maria D, Gupta N, Stern J, O'Connor K, Hafler D, Laserson U, Vigneault F, Kleinstein S. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol. 2013;4:358. https://doi.org/10.3389/fimmu.2013.00358.
56. Spisak N, Walczak AM, Mora T. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. Nucleic Acids Res. 2020;48(19):10702–10712.
57. Zhang B, Meng W, Prak ETL, Hershberg U. Discrimination of germline v genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment. J Immunol Methods. 2015;427:105–116.
58. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnol. 2017;17(1):1–16.
59. Lynch JN, Donermeyer DL, Weber KS, Kranz DM, Allen PM. Subtle changes in TCRα CDR1 profoundly increase the sensitivity of CD4 T cells. Mol Immunol. 2013;53(3):283–294.
60. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, Özkan E, Davis MM, Wucherpfennig KW, Garcia KC. Deconstructing the peptide-MHC specificity of T cell recognition. Cell. 2014;157(5):1073–1087.
61. Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biol. 2017;18(1):1–21.
62. Ralph DK, Matsen IV FA. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. PLoS Comput Biol. 2016;12(1):1004409.
63. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. Mol Immunol. 2017;87:12–22.
64. Yu Y, Ceredig R, Seoighe C. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. Nucleic Acids Res. 2016;44(4):31–31.
65. Schwartz GW, Hershberg U. Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. Front Immunol. 2013;4:357.
66. Shin EK, Matsuda F, Ozaki S, Kumagai S-i, Olerup O, Ström H, Melchers I, Honjo T. Polymorphism of the human immunoglobulin variable region segment v1-4.1. Immunogenetics. 1993;38(4):304–306.
67. Chimge N, Pramanik S, Hu G, Lin Y, Gao R, Shen L, Li H. Determination of gene organization in the human IGHV region on single chromosomes. Genes Immun. 2005;6(3):186–193.
68. Rubinstein DB, Symann M, Stewart AK, Guillaume T. Restriction fragment length polymorphisms and single germline coding region sequence in VH182, a duplicated gene encoding autoantibody. Mol Immunol. 1993;30(4):403–412.
69. Shin E, Matsuda F, Nagaoka H, Fukita Y, Imai T, Yokoyama K, Soeda E, Honjo T. Physical map of the 3' region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. EMBO J. 1991;10(12):3641–3645.
70. Sasso E, Buckner J, Suzuki L. Ethnic differences in VH gene polymorphism. Ann N Y Acad Sci. 1995;764(1):72–73.
71. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol. 2010;184(12):6986–6992.
72. Peng K, Safonova Y, Shugay M, Popejoy AB, Rodriguez OL, Breden F, Brodin P, Burkhardt AM, Bustamante C, Cao-Lormeau V-M, et al. Diversity in immunogenomics: the value and the challenge. Nat Methods. 2021;18(6):588–591.
73. Slabodkin A, Chernigovskaya M, Mikocziova I, Akbar R, Scheffer L, Pavlović M, Bashour H, Snapkov I, Mehta BB, Weber CR, et al. Individualized VDJ recombination predisposes the available Ig sequence space. bioRxiv. 2021.
74. Chung W, Hung S, Hong H, et al. Medical genetics: a markerfor Stevens-Johnson syndrome. Nature. 2004;428:486.
75. Pan R-Y, Dao R-L, Hung S-I, Chung W-H. Pharmacogenomic advances in the prediction and prevention of cutaneous idiosyncratic drug reactions. Clin Pharmacol Ther. 2017;102(1):86–97.
76. Pan R-Y, Chu M-T, Wang C-W, Lee Y-S, Lemonnier F, Michels AW, Schutte R, Ostrov DA, Chen C-B, Phillips EJ, et al. Identification of drug-specific public tcr driving severe cutaneous adverse reactions. Nat Commun. 2019;10(1):1–13.
77. Lin M-J, Lin Y-C, Chen N-C, Luo AC, Lai S-K, Hsu C-L, Hsu JS, Chen C-Y, Yang W-S, Chen P-L. Profiling germline adaptive immune receptor repertoire with gAIRR suite. bioRxiv. 2020.
78. Omer A, Peres A, Rodriguez OL, Watson CT, Lees W, Polak P, Collins AM, Yaari G. T cell receptor beta germline variability is revealed by inference from repertoire data. Zenodo. 2021. https://doi.org/10.5281/zenodo.5652127. https://doi.org/10.5281/zenodo.5652127.

79. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, Boyd SD, Collins AM, Lees W, Yaari G. VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res. 2020;48(D1): 1051–1056.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.