T Cell Receptor Beta (TRB) 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]

[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, United States
[4]Institute of Structural and Molecular Biology, Birkbeck College, University of London, London, United Kingdom
[5]School of Biotechnology and Biomedical Sciences, University of New South Wales, Sydney, Australia
[*]These authors equally contributed to this work.
[†]To whom correspondence should be addressed. Tel: +972 3 7384625; Email: gur.yaari@biu.ac.il

May 17, 2021

1 Abstract

T and B cell 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 T and B cell receptors (TCRs and BCRs, respectively) are challenging to decipher due to repetitive elements and undocumented structural variants. To confront this challenge, AIRR-seq-based methods have been developed recently for B cells, enabling genotype and haplotype inference and discovery of undocumented alleles. Applying these methods to AIRR-seq data reveals a plethora of undocumented genomic variations. However, this approach relies on complete coverage of the receptors' variable regions, and most T cell studies sequence only a small fraction of the variable region. Here, we adapted BCR inference methods to full and partial TCR sequences, and identified 38 undocumented polymorphisms in TRBV, 15 of them were also observed in genomic data assemblies. Further, we identified 31 undocumented 5' UTR sequences. A subset of these inferences was also observed using independent genomic approaches. We found the two documented TRBD2 alleles to be equally abundant in the population, and show that the single nucleotide that differentiates them is strongly associated with dramatic changes in the expressed repertoire. Our findings expand the knowledge of genomic variation in the TRB (T Cell Receptor Beta) locus and provide a basis for annotation of TCR repertoires for future basic and clinical studies.

2 Introduction

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 [21]. It can teach us about fundamental immune processes and reveal dysregulation, with broad implications for biomedicine. B and T cell receptors are assembled during differentiation of hematopoietic stem cells, by a complex process involving somatic recombination of a large number of germline-encoded V, D, and J gene segments, along with junctional diversity that is added at the boundaries where these segments are joined together [40]. 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 [23, 52]. 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 [9]. 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 [25]. Complete sequences of the TCR encoding loci were reported in 1985 [24, 70, 61], and this led to the development of the TCR encoding genes nomenclature by the ImMunoGeneTics (IMGT) group [30]. 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 may be far from complete [33, 35, 60].

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 [66, 16]. This changed, both with a method for targeted long-read direct sequencing of the BCR heavy chain encoding locus [49], and with the realisation that BCR encoding undocumented germline alleles and genotypes can be reliably inferred from AIRR-seq data [10, 19, 17, 47], as well as haplotypes [28, 45], and chromosomal deletions within the BCR encoding loci [22]. 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 [72, 27, 33], but these approaches were not validated with targeted sequencing or AIRR-seq, and therefore are subject to criticism regarding the reliability of the inferences [65, 9]. 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 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 [62] amplify partial VDJ sequences from the J gene to the framework-2 (FR2) of the V gene, while the Adaptive Biotechnologies [48] approach generates only 87 nucleotides from a fixed position within each 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, it can be impossible to

unambiguously identify the gene and allele source of these partial V(D)J sequences. Thus, the development of a TRBV genotype inference method requires the thorough 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, 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). This is an even greater problem in 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. These results may have important implications to TCR based immunotherapy and disease diagnosis.

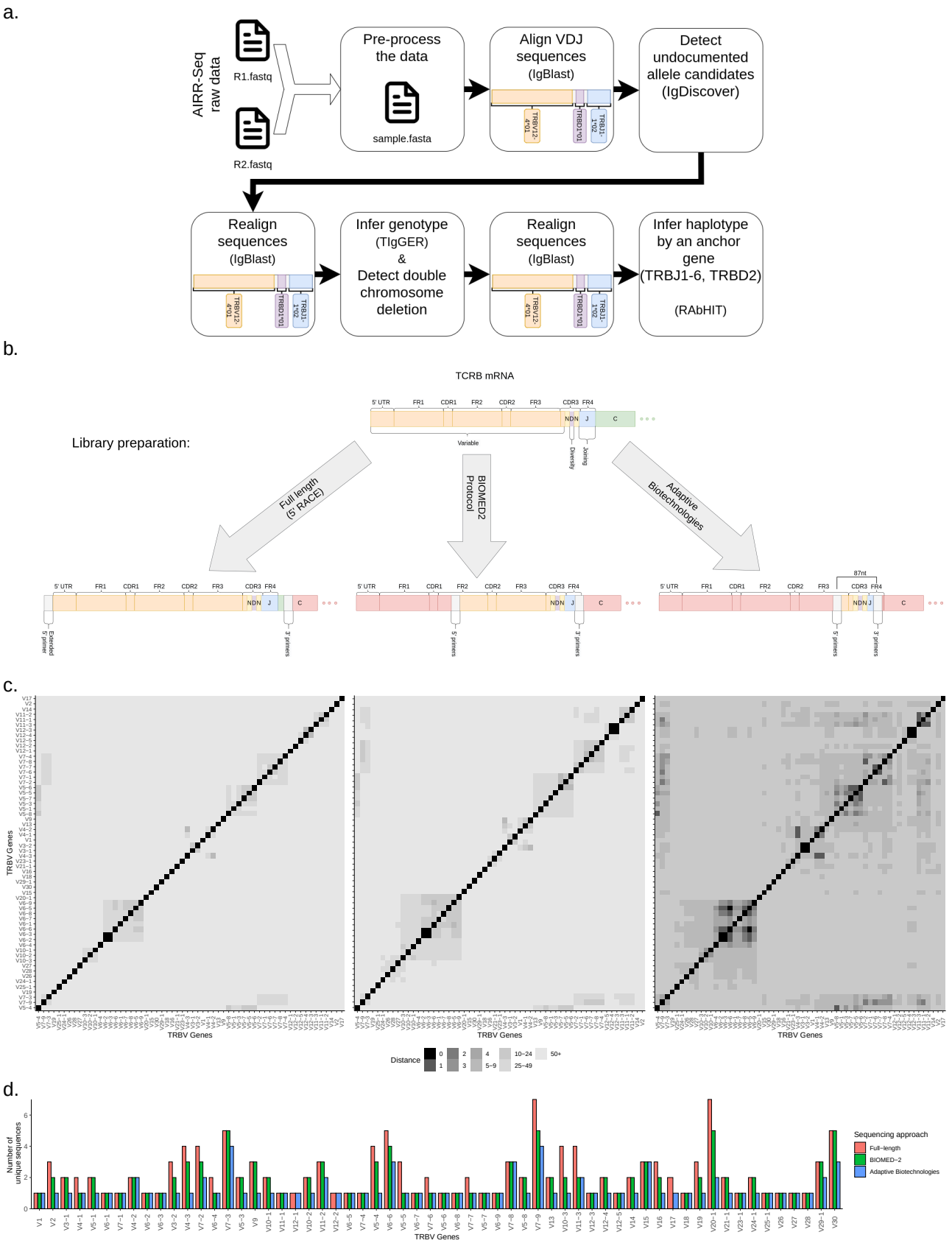

Figure 1: **Overview of TRB genotyping and haplotyping workflow, TCRβ sequencing technologies, and number of documented alleles.** (a) An illustration of the steps for TRB genotyping and haplotyping. (b) An illustration of the three TCRβ sequencing approaches investigated in this study. (c) Heatmaps of distances between amplified TRBV gene sequences. The three panels corresponds to the sequencing approaches shown above them. Colors correspond to the distances between the genes' as measured by the Hamming distance between the closest two alleles in the sequenced region. (d) Barplot of the unique sequence number per gene in the the studied sequencing approaches.

3 Results

3.1 Identification of allelic variation 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 (personal genotypes). For this, we took a multi step approach, in which we first assigned each sequence to the most likely allele(s) using IgBLAST [69] with a reference panel of all the documented TRBV alleles as they appear in IMGT [30]. Then, we used IgDiscover [10] to detect candidate sequences that have not previously been documented in IMGT. Hereafter, we will refer to such sequences as "undocumented alleles". Following this step, we re-ran IgBLAST [69] with an extended reference germline sequence set that includes the candidates for undocumented alleles. Finally, we inferred the personal genotype of each gene and individual using TIgGER [20]. To ensure our confidence in the inference of undocumented alleles, we took steps that reduce the influence of sequencing errors. First, we inspected the allele usage within the data-set of each study population. We expect the population distribution of usage to be either bi or tri-modal, as a result of usage in individuals who are either heterozygous or homozygous for the allele, as well as individuals who do not carry the allele. Second, we inspected the adjacent nucleotides to each single nucleotide polymorphism (SNP) position. No candidate SNP within a stretch of four or more identical nucleotides were considered, as such a stretch is more likely to arise from a sequencing read artefact.

We applied the above approach to four data-sets, spanning different sequencing protocols and scales (see methods). From 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 uni-modal 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 adjacent 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 (Supp. Table 5). The first, a shotgun sequencing analysis of 286 individuals from Luo et al. [33], where six out of the 12 undocumented alleles were observed. The second was a pmDb [27] data-set, where we found four out of the 12 undocumented alleles. The third was long-read assemblies of 35 diploid from Ebert et al. [13] which confirmed six out of the 12 undocumented alleles. All together, two of the 12 undocumented alleles were observed in all sources.

In addition to the 18 undocumented alleles, five additional undocumented allelic variants were identified in DS1 that matched known allele references in IMGT, except in their 3' end that is missing in IMGT (Supp. Table 3). We replaced these IMGT alleles with their longer versions in our analyses, since it improves the genotype inference.

Another genomic variation we explored in DS1 is in the 5' UTR. Consensuses of the 5' UTR were constructed as previously described in Mikocziova et al. [38, 39]. 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 (Supp. Table 1), 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 (Supp. Table 2). Four out of the 11 L-PART1 absent sequences and six out of the 10 L-PART2 absent sequences were also observed in long-read assemblies (Supp. Table 2). Two sequence

variations 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 variations from 14 different alleles, 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 [36]. 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 (Supp. Fig. 7). 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 (Genome Reference Consortium Human Build 38) [55]. 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 misidentified in IMGT (Supp. Fig. 8). We also found three variant consensus sequences associated with TRBV6-2*01/TRBV6-3*01 (Supp. Fig. 7), 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. Five others were present in the short-read whole genome databases, one in Luo et al. [33] and four in pmDB [27]. Two out of those five undocumented alleles were observed in the long-read assemblies of 35 diploid from Ebert et al. [13]. One out of those two undocumented alleles was observed in all sources (Supp. Table 5).

Lastly, we investigated the 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. From the 313 DS3 TRB genotypes [58], we inferred 27 potential undocumented allele patterns. However, the usage fraction of 12 of them did not follow a bi-modal or a tri-modal distribution (Supp. Fig. 5). Two more undocumented alleles had SNPs in positions within stretches of identical nucleotides. We therefore discarded those 14 unreliable inferences. Two out of the remaining 13 undocumented alleles were independently identified in multiple genotypes, giving us added confidence in these inferences (Supp. Fig. 5). Two alleles were supported by the presence of alleles identified also by Luo et al. [33]. Two alleles were supported also by pmDB [27]. A similar analysis of inferred genotypes from 786 individuals in the Adaptive Biotechnologies data-set (DS4) identified a further 18 undocumented alleles (Supp. Fig.6), four were discarded for not following the bi-modal distribution and one more for an SNPs in positions within stretches of identical nucleotides. To summarize this part, we compared the documented alleles for this locus versus all alleles that were observed in DS1 and DS2, DS3, DS4, and to the assemblies of long reads from genomic data (Fig. 2b). All in all we identified 38 undocumented alleles. Out of those 15 alleles were also observed in the long-read assemblies of 35 diploid from Ebert et al. [13].

3.2 Double chromosome deletions

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 discussion). TRBV gene usage in DS1 shows such deletion poly-

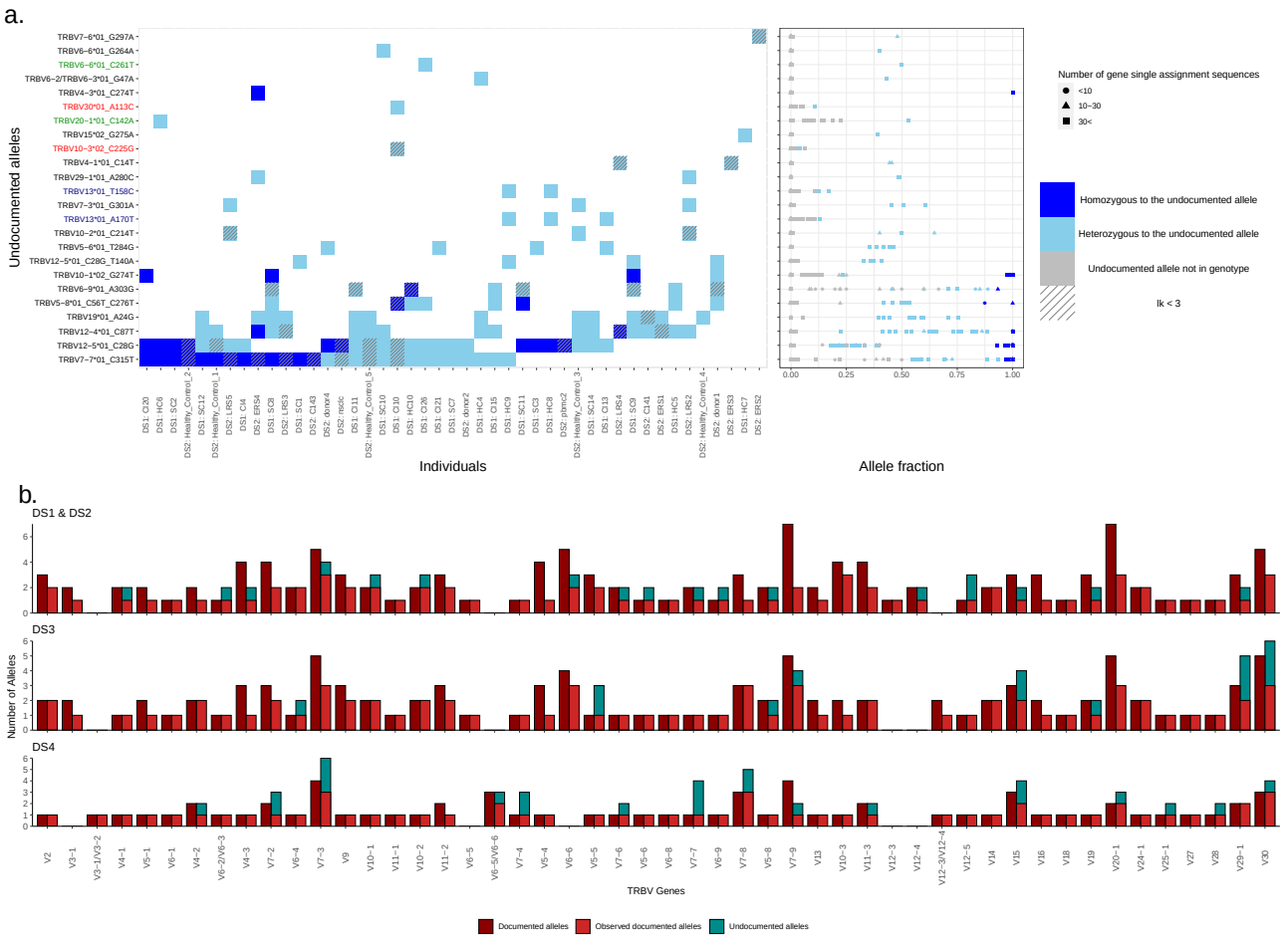

Figure 2: **Undocumented alleles.** (a) Undocumented allele distribution in DS1 and DS2. The left panel shows undocumented alleles within the genotype. Each row represents an undocumented allele, each column represents an individual. The Y-axis annotations in red, green, and blue correspond to alleles not following the expected multi-modal distribution, alleles 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.

morphisms in four genes (Fig. 3a). TRBV4-3 and TRBV3-2 are absent from the genotypes of eight individuals, 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 (Supp. Fig. 3). 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 $\sim$21Kb deletion which includes TRBV4-3, TRBV3-2 and TRBV6-2 has been reported from genomic studies [73, 51, 6, 30]. The inference of TRBV6-2 deletions in AIRR-Seq data is made difficult because of the existence of TRBV6-3, which is an exact duplicate of TRBV6-2. However, indirect evidence that the deletion polymorphisms seen in DS1 are associated with the previously-reported $\sim$21Kb 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 (Supp. Fig. 13). It is therefore likely that detection of TRBV6-2*01/TRBV6-3*01 in these individuals is entirely a consequence of sequences utilizing TRBV6-3*01. This line of reasoning also allowed us to conclude that an undocumented polymorphism seen in sample HC4 is most likely an allele of TRBV6-3 (TRBV6-3*01_G47A) rather than TRBV6-2. The genotypes of all individuals who carry the TRBV6-2*01/TRBV6-3*01_2 or TRBV6-3*01_G47A 5'UTR sequences, include TRBV7-2*02. Since the presence of TRBV7-2*02 hints at deletion of TRBV6-2, we conclude that the TRBV6-2*01/TRBV6-3*01_2 5'UTR sequences is attributed to TRBV6-3*01.

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).

3.3 D and J genotyping and their association

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. During VDJ 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 [41]. Studies of N-nucleotide addition in BCR V(D)J genes shows the process to be biased towards addition of Gs, and addition of homopolymer tracts [26]. 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

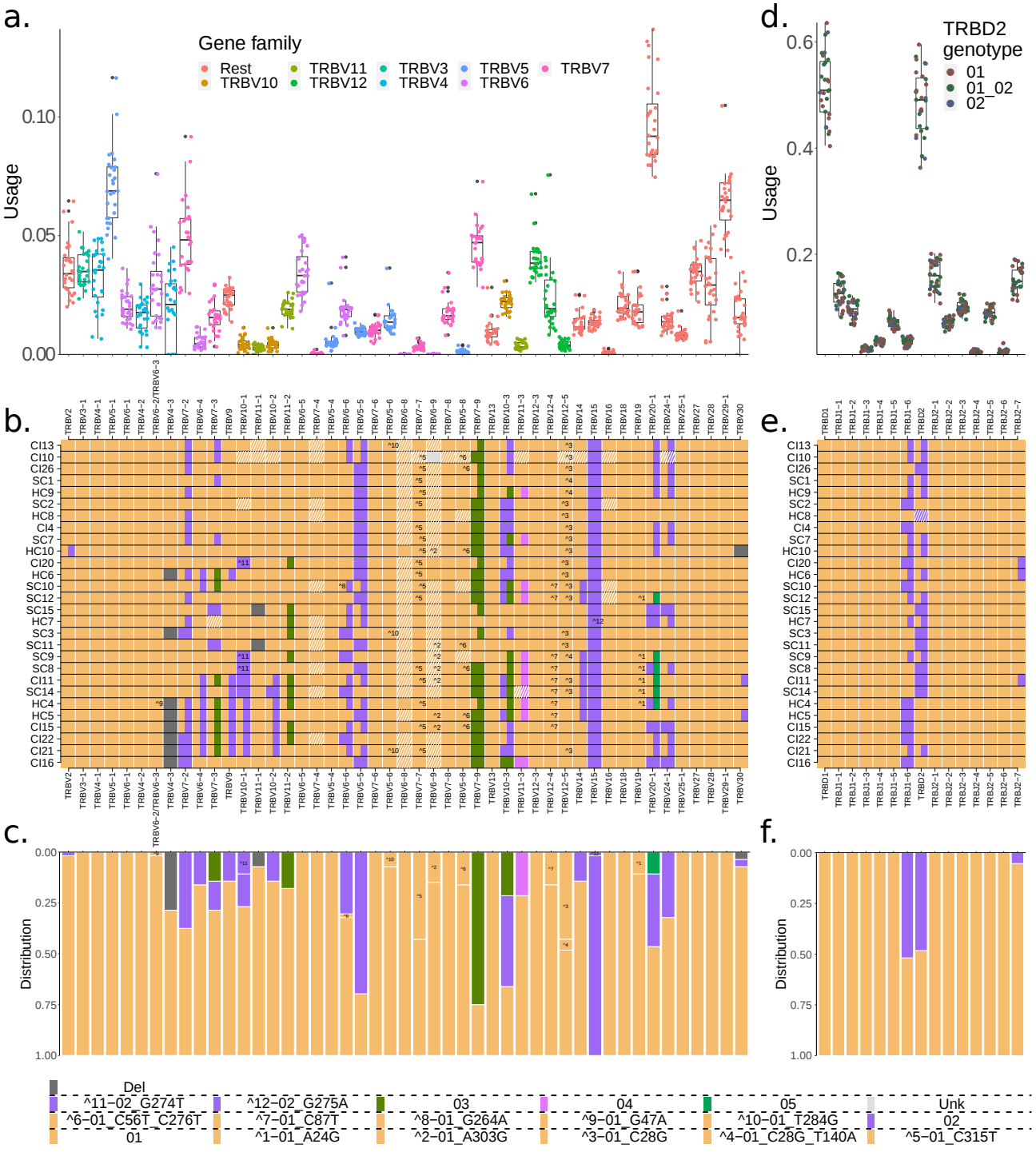

Figure 3: **Gene usage and genotypes of the DS1 data-set.** (a) TRBV gene usage. The X-axis shows the TRBV genes in the order in which they are found in the genome. The Y-axis shows the frequency of gene usage, colored by gene family. (b) TRBV genotypes. Each row shows an individual genotype, and columns correspond to the different genes in the order in which they are found in the genome. Colors correspond to alleles as indicated in the bottom of the figure. (c) TRBV allele frequency. The X-axis shows the TRBV genes in the order in which they are found in the genome, and the Y-axis shows the frequency of the alleles in the population, colored by allele. Panels (d-f) show analogous plots to (a-c) but for the TRBD and TRBJ genes. Colors in (d) correspond to the TRBD2 genotype of the individual.

errors. 23 out of the 28 individuals from DS1 were initially inferred to be heterozygous at the TRBD locus. However, this is most 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 [2]. To document TRBD2 misassignments, we examined the fraction of TRBD2*01 assignments amongst all sequences unambiguously assigned to the TRBD2 gene. The observed frequencies defined three very distinct groups (Fig. 4a). In the first group, TRBD2*01 assignments accounted for $\sim$0.125 of the TRBD2 annotations, in the second group 0.3-0.6, and in the third group $\sim$0.95. We believe that individuals in the first group are in fact 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 the third group is homozygous for TRBD2*01. The resulting frequency distribution of TRBD2 fits well the Hardy–Weinberg principle.

To better define the thresholds differentiating the three groups, we turned to the larger data-sets. DS3 contains 348 samples from 313 individuals, however, the distribution of TRBD2*01 frequencies was too noisy (Fig. 4b), 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 (Fig. 4c). The homozygous TRBD2*01 group is centered around a frequency of 0.96. The heterozygous group is centred around 0.45, and the homozygous TRBD2*02 group is centered around 0.12. The medians of the three groups were close to their averages (Table. 1), thus for the current purpose we could assume that the three data-sets are normally distributed.

TRBD2 genotype group \ TRBD2*01 fraction	Average	Median	Standard deviation
TRBD2*01 homozygous	0.96	0.96	0.003
TRBD2 heterozygous	0.453	0.452	0.022
TRBD2*02 homozygous	0.127	0.127	0.008

Table 1: The distribution parameters of the TRBD2*01 fraction according to the TRBD2 genotype group in DS4.

We determined that the borders between groups are at the points where the TRBD2*01 fraction has an equal probability to be in either one of the two adjacent groups. These equilibrium points were calculated as described in the methods. This 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 0.2066-0.8968, and the homozygous TRBD2*01 group was composed of 190 samples with a TRBD2*01 fraction above 0.8968. Thus, the D2 allele frequency was around 0.244 homozygous individuals for TRBD2*01, 0.514 heterozygous individuals, and 0.242 homozygous individuals for D2*02. As mentioned, these allele frequencies accord with the Hardy-Weinberg principle.

These findings imply 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. This strong bias may result from selection resulting from the amino acid differences in the sequences or from unknown structural variations in the locus that could be associated with the different alleles.

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*01 is the only TRBJ gene with two known functional alleles.

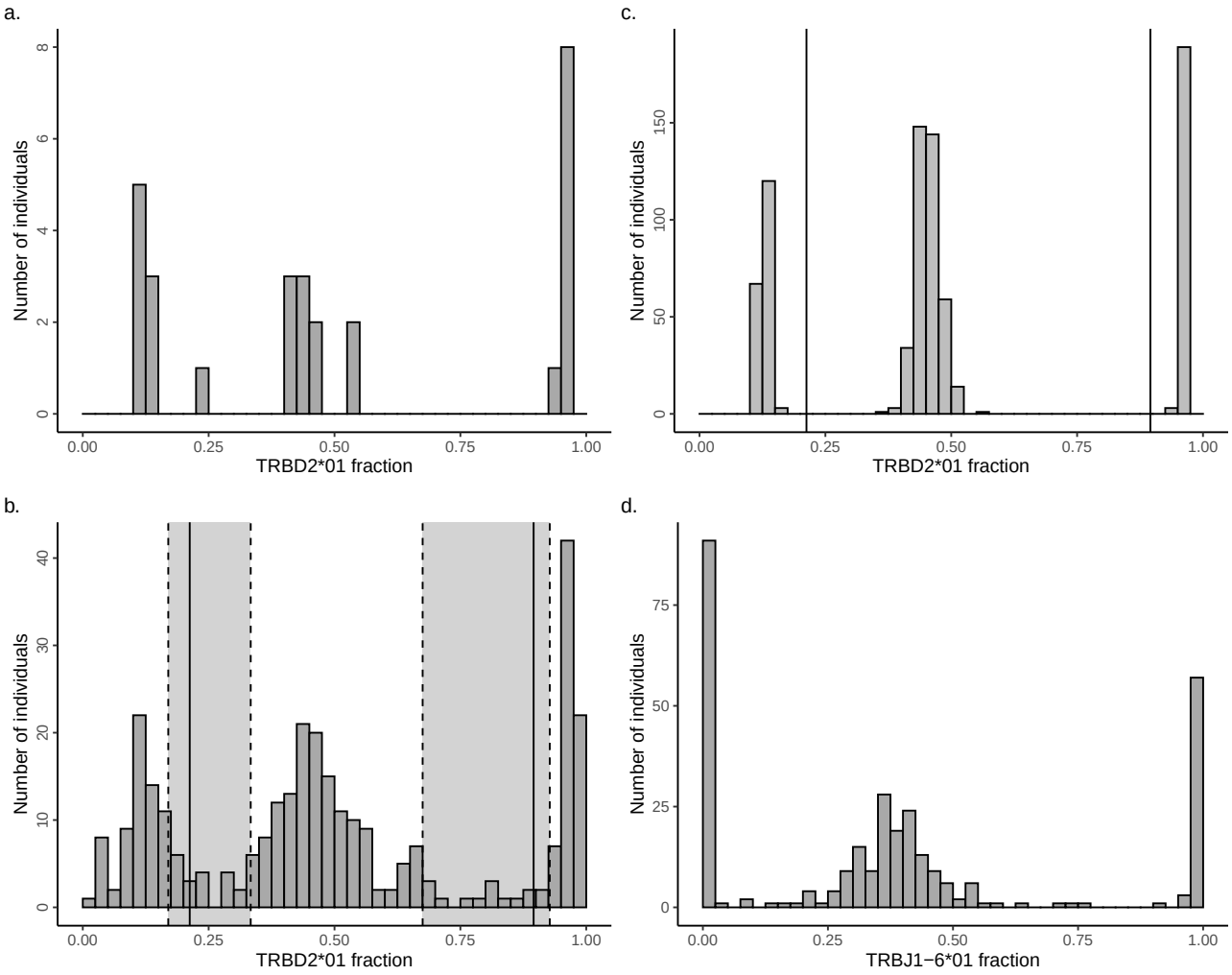

Figure 4: **TRBD2 and TRBJ1-6 genotypes and gene usage frequencies.** (a) The frequency of TRBD2*01 usage, as a fraction of total D2 usage in 28 individuals from the DS1 data-set. Homozygous and heterozygous genotypes can be inferred from the distribution. (b) The frequency of TRBD2*01 usage, as a fraction of total D2 usage in 313 individuals from the DS3 data-set. Solid vertical lines indicate boundaries between homozygous and heterozygous individuals, as calculated from analysis of DS4 data-set. The genotypes of individuals whose frequencies fall within the shaded regions cannot be inferred with confidence. (c) The frequency of TRBD2*01 usage, as a fraction of Total TRBD2 usage in 786 individuals from the DS4 data-set. Solid vertical lines indicate boundaries between homozygous and heterozygous individuals. (d) The frequency of TRBJ1-6*01 usage, as a fraction of Total TRBJ1-6 usage in 313 individuals from the DS3 data-set.

The distribution of frequencies shows a good partitioning between homozygous and heterozygous individuals (Fig. 4d), 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 (Fig. 4d). 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 ($\sim$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 (Supp. Fig. 12). We observed SNPs with a high LD score between the genes. Further, 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 Fig. 4b. Supp. Fig. 9 shows that all of the homozygous TRBD2*01 individuals are also homozygous for TRBJ1-6*02. Also, 50 out of the 52 homozygous TRBJ1-6*01 individuals are homozygous for TRBD2*02.

3.4 TRBD and TRBJ usage

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 1.25 times more than homozygous TRBD2*01 individuals (Fig. 5a, left panel). TRBD2 can undergo rearrangements only with TRBJ2 genes [64], 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. 5a, 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. 5b).

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. 5c). 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. 5d shows the conditional probability P(TRBJ2-N|TRBD2) for all the TRBJ2 genes. The biased usage of the TRBJ2 genes observed in Fig. 5c 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. 5e), and observed that TRBD2 genotype is associated with the TRBD1 likelihoods to re-

arrange 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 P(TRBJ2-N|TRBD1), Fig. 5f).

The strong biases observed in Fig.5a-e can result from amino acid alterations in the sequence, from non-coding regulatory variants, or from unknown structural variations 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 thymus selection [12, 68, 59]. In these non-functional sequences the biases are pronounced in a similar fashion (Supp. Fig. 11). Thus, we conclude that the differences between the TRBD and TRBJ rearrangements stratified by D2 genotype are most likely due to structural differences or non-coding regulatory variants between the loci rather than due to negative selection.

3.5 Haplotype inference reveals association patterns

From the genotype analysis of DS1, we observed a potential pattern between homozygousity 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 [28, 45], 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 3e) and 8 individuals from DS2 (Supp. Fig. 3) are heterozygous. In 7 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, CI21_J1-6_01, CI21_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 strengthen this effect, we surveyed individual genotype and haplotype inferences in DS3. DS3 is a much larger data-set, which 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 (Supp. Table 1 BIOMED-2). In 11 out of 12 individuals with a high genotype inference likelihood, the pattern between homozygosity of TRBV7-2*bp02 (see section 4.3) and a deletion inference of TRBV4-3 was apparent (Supp. Fig. 14 and 15). 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 TRBV6-2 and TRBV4-3, are not present in haplotypes that carry TRBV7-2*bp02.

In addition, the following association patterns between specific alleles and single chromosome deletions were revealed by the haplotype inference: 1. TRBV24-1*02 was observed in 8 samples on the chromosome carrying TRBJ1-6*02 and none in the other chromosome. 2. TRBV29*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 5 samples on the chromosome carrying TRBJ1-6*02 and none in the other chromosome. 4. In 8 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. 6).

TRBD2 was also considered a potential anchor gene for haplotype inference. To examine

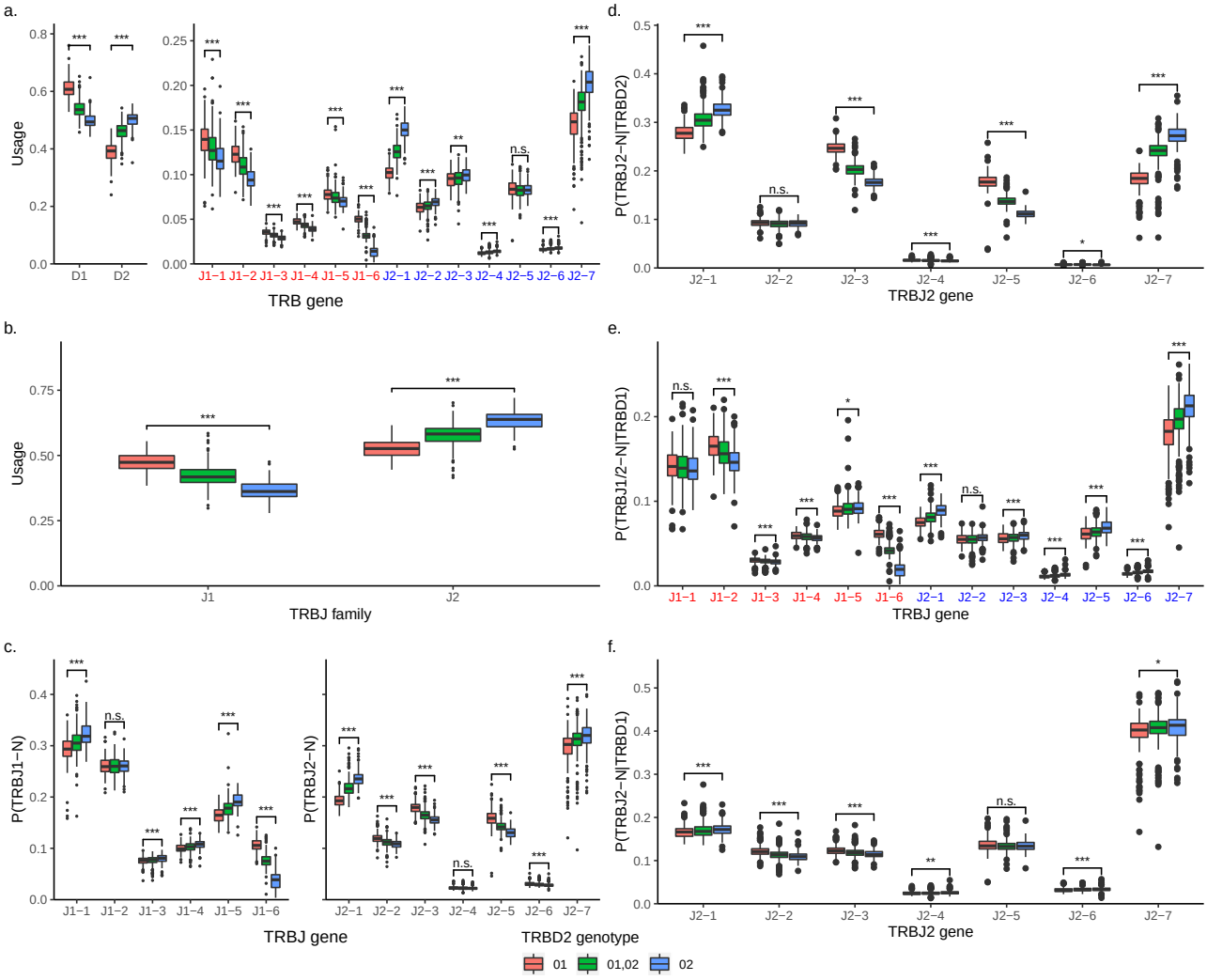

Figure 5: **TRBD and TRBJ usage correspond to TRBD2 genotype.** (a) TRBD and TRBJ gene usage in DS4 individuals with different TRBD2 genotypes. TRBJ genes are shown along the X-axis in the order in which they appear in the genome. (b) The TRBJ family usage in DS4 individuals with different TRBD2 genotypes. 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 *** - $p < 0.001$).

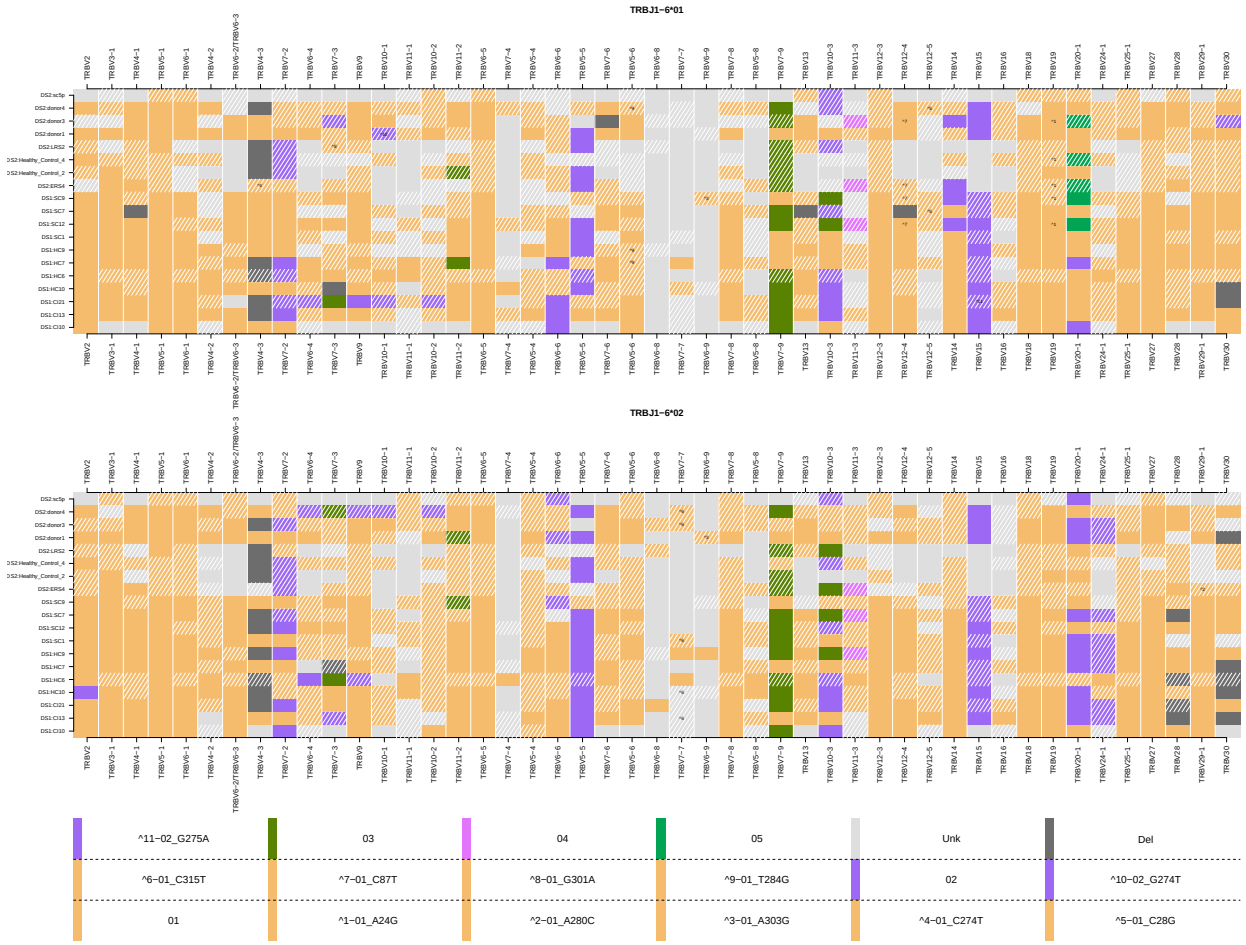

Figure 6: TRBV haplotypes for 19 individuals from DS1 and DS2. The upper panels show the TRBV haplotypes anchored with TRBJ1-6*01 and TRBJ1-6*02, respectively. Each row is an individual's haplotype, and each column is a V gene call. The colors corresponds to the V alleles and the tile annotations correspond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

the reliability of this inference, we used the same individuals who were heterozygous for TRBJ1-6. 10 out the 11 individuals from DS1 were heterozygous for TRBD2 and were used for the haplotype inference (Supp. Fig. 16). 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.

4 Methods

4.1 Data

Three types of TCRβ AIRR-seq data-sets were collected [14, 58, 15]. Each type was produced using a different sequencing protocol, and each protocol produces sequences of different lengths. An additional data-set of TCRs amplified from single-cells was collected from three different sources. Data came from 14 individuals from a COVID-19 study, and was accessed via iReceptor plus [11, 67]. Additional data came from three individuals from a second COVID-19 study [31]. Data from 8 individuals was provided by 10x Genomics as shown in Supp. Table 9. The data-sets are described in Table. 2. DS2 and DS4 were downloaded after preprocessing, DS1 was preprocessed according to the preprocessing of Eliyahu et al. [14], and DS3 was preprocessed using pRESTO [63] 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, (iv) duplicate sequences were removed and collapsed.

Data-set	**Cohort**	**# of Individuals**	**# of Samples**	**Sequencing protocol**	**UMI**	**Helix**	**Reference**
DS1	HCV	28	28	5' RACE (full-length)	+	RNA	PRJEB28370
DS2	-	25	25	10x Genomics (full-length)	+	RNA	[1], PRJNA608742, HRA000069
DS3	Cancer	313	348	BIOMED-2	-	DNA	PRJEB33490
DS4	CMV	786	786	Adaptive Biotechnologies	-	DNA	[3]

Table 2: **Data-sets analyzed in this study.**

4.2 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 partial genes 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 [48]. 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. 96% of the sequences cover the first position following the TRBV gene primers (Supp. Fig. 2). 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.

4.3 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 Supplementary Table 8 and Supplementary Table 7) Sequences that could not be identified to the allele level are reported here without reference to allele numbers. For example, where the sequences are identified in the DS4 study, all four known alleles of the TRBV4-3 gene are reported as TRBV4-3. The primers of the BIOMED-2 protocol were taken from van Dongen et al. [62], and the primers of Adaptive Biotechnologies were taken from Robins et al. [48].

4.4 Genotype and undocumented allele inference

IgDiscover [10] 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. TRBV allele candidates were inferred using the discover function from the same software. 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 assignments. 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.

Genotyping involved the following steps. Pre-processed FASTA files were aligned against the IMGT TRBV, TRBD, and TRBJ Reference Directories using IgBLAST [69]. Putative new alleles were inferred for each individual using IgBLAST, and 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 [18]. Genotyping was limited to sequences with a single assignment (only one best match), and without mismatches 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 9 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.

4.5 Validating undocumented alleles/variants in long-read assemblies

35 diploid (70 haplotypes) HIFI long-read assemblies were downloaded from Ebert et al. [13] and aligned to GRCh38 [55] using BLASR with default parameters. 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.

4.6 Calculation of the equilibrium point between two normal distributions

To refine TRBD genotypes, resolving errors resulting from misidentification of TRBD2 genes, it was necessary to identify boundaries to separate individuals who are homozygous at the TRBD2 locus from individuals who are heterozygous at the locus. Frequency distributions of TRBD2*01 usage levels in different data-sets were analysed and the the nature of the distributions within the tri-modal peaks were considered. Boundaries between the 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:

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

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

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

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

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

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

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

4.7 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.

4.8 Double chromosome deletion inference

The detection of double chromosome deletion was done using the published method [22], with a single difference that the minimum cutoff of the average gene usage was lowered from 0.005 to 0.0005. A data frame table that contains the following columns: individual, gene,

N, and total was used. 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.

4.9 Haplotype inference

RAbHIT was used as previously described [45], 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 Fig. 4) mentioned above. As $\sim$12.5% of TRBD2*02 rearrangements mis-align to TRBD2*01, epsilon was set at 0.125 if this allele 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.

5 Discussion

Despite the intimate partnership of the Major Histocompatibility Complex (MHC) and TCRs in the recognition of antigenic peptides, it is only the genes of the MHC that are widely recognised as disease susceptibility genes [37]. 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 what 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 the TCR repertoire now provide a new avenue by which germline genes can be easily documented, for this study has shown that tools and techniques that were developed for the analysis of individual BCR genotypes - the sets of BCR encoding genes that are available for rearrangement [22, 19, 28, 46, 17, 29], and individual haplotypes - the available chromosomal sets of IGHV genes [10, 45], can easily be adapted for the analysis of TCR data-sets. Other tools such as LymAnalyzer have been specifically developed for the analysis of TCR data [71].

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 previously been extended to investigations of the TCR genes. The focus of most TCR gene studies remains firmly fixed on the CDR3 regions of the genes [50]. Despite their interaction with the MHC, and their documented influences on TCR/MHC/peptide interactions [34, 4], 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. In this study, we have explored how TCR VDJ sequences of different lengths, generated using differing technologies, can contribute to our knowledge of TRB genes and their allelic variants.

Our analysis demonstrates that many undocumented V genes remain to be discovered, and that even short read sequences can be analysed for the detection of undocumented polymorphisms. It is clear, however, that much additional information is to be gained by

the study of full length VDJ 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 13 polymorphisms were identified in truncated sequences generated from 313 individuals using BIOMED-2 primers, and just 16 new SNPs were seen in the 786 Adaptive Biotechnology data-sets. This latter result reflects both the very short lengths of the Adaptive Biotechnology sequences, and the general lack of variability in the 3' ends of the TRBV genes.

It is interesting to contrast this level of gene discovery with genetic variation amongst the BCR-encoding genes: to date, 64 functional alleles are documented in IMGT for 48 TRBV genes (avg. 1.33 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 38 alleles and SNPs to the record of TRBV genes, 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 21Kb 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 - TRBV4-1, TRBV4-2, TRBV7-3, TRBV28 and TRBV30 - were seen at relatively low frequencies. No deletion polymorphisms involving the TRBD or TRBJ loci were detected. In contrast, numerous relatively common indels are now known in the IGH loci, including one involving 13 consecutive, functional IGHV genes, and one involving 6 consecutive IGHD genes [28, 22]. 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 indel patterns [22, 49].

It should be emphasised 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 individual variations 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 8 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 heavy chain IGHV locus. Early RFLP-based analyses pointed to duplications of sequences that came to be known as IGHV3-23 and IGHV1-69 [57, 7, 53, 56, 54]. More recently, the duplication of these two genes and of three other functional genes were confirmed, first by analysis of AIRR-Seq data [5] and then with the publication of a second complete assembly of the IGHV locus in 2013 [66].

It was not possible in this study to explore possible RSS variants, as RSS are lost from the genome during VDJ recombination. On the other hand, 5' RACE data allows for the exploration of variations in the 5' UTR, as has recently been reported from BCR repertoire studies [38, 39]. 31 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 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 [23, 52], 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 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 neighbouring 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 variations between the TRBD genes have consequences for the expressed repertoire. Within large VDJ 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. 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.

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 [8]. SJS and other cutaneous hypersensitivity reactions have been linked to HLA types, but these associations have all had a low positive predictive value [43], 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 [44]. Full length sequences were not reported in the study of Pan and colleagues [44], 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. SNP coverage of these regions is even sparser than coverage of the BCR gene loci, and therefore real associations will escape detection in genome-wide association studies. 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 [32], 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 expand our knowledge of genomic variations in the TRB locus and lay the ground for more accurate basic and clinical future studies.

TCR repertoire generative models such as IGoR do not take germline variation explicitly into account. As seen in Figure 4/5 this has a huge effect on the repertoire.

Data Availability

Alleles inferred in this study are listed in Supplementary Data. Repertoire analyses of the four data-sets will be published in VDJbase [42] (https://www.vdjbase.org) at the time of publication.

Acknowledgements

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.

References

[1] 10x Genomics. 10x datasets. https://support.10xgenomics.com/single-cell-vdj/datasets. Accessed: 2020-12-15.

[2] C Andrews. The hardy-weinberg principle. *Nature Education Knowledge*, 3(10):65, 2010.

[3] Adaptive Biotechnologies. Adaptive biotechnologies datasets. https://clients.adaptivebiotech.com/pub/Emerson-2017-NatGen. Accessed: 2020-12-15.

[4] Michael E Birnbaum, Juan L Mendoza, Dhruv K Sethi, Shen Dong, Jacob Glanville, Jessica Dobbins, Engin Özkan, Mark M Davis, Kai W Wucherpfennig, and K Christopher Garcia. Deconstructing the peptide-mhc specificity of t cell recognition. *Cell*, 157(5):1073–1087, 2014.

[5] Scott D Boyd, Bruno A Gaëta, Katherine J Jackson, Andrew Z Fire, Eleanor L Marshall, Jason D Merker, Jay M Maniar, Lyndon N Zhang, Bita Sahaf, Carol D Jones, et al. Individual variation in the germline ig gene repertoire inferred from variable region gene rearrangements. *The Journal of Immunology*, 184(12):6986–6992, 2010.

[6] Rebekah M. Brennan, Jan Petersen, Michelle A. Neller, John J. Miles, Jacqueline M. Burrows, Corey Smith, James McCluskey, Rajiv Khanna, Jamie Rossjohn, and Scott R. Burrows. The impact of a large and frequent deletion in the human tcrβ locus on antiviral immunity. *The Journal of Immunology*, 188(6):2742–2748, 2012.

[7] NO Chimge, S Pramanik, G Hu, Y Lin, R Gao, L Shen, and Ho Li. Determination of gene organization in the human ighv region on single chromosomes. *Genes & Immunity*, 6(3):186–193, 2005.

[8] WH Chung, SI Hung, HS Hong, et al. Medical genetics: a markerfor stevens-johnsonsyndrome. *Nature*, 428:486, 2004.

[9] Andrew M Collins, Gur Yaari, Adrian J Shepherd, William Lees, and Corey T Watson. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? *Current Opinion in Systems Biology*, 2020.

[10] Martin M. Corcoran, Ganesh E. Phad, Néstor Vázquez Bernat, Christiane Stahl-Hennig, Noriyuki Sumida, Mats A.A. Persson, Marcel Martin, and Gunilla B. Karlsson Hedestam. Production of individualized v gene databases reveals high levels of immunoglobulin genetic diversity. *Nature Communications*, 7(1):13642, Dec 2016.

[11] B. D. Corrie, N. Marthandan, B. Zimonja, J. Jaglale, Y. Zhou, E. Barr, N. Knoetze, F. M. W. Breden, S. Christley, J. K. Scott, L. G. Cowell, and F. Breden. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev*, 284(1):24–41, 07 2018.

[12] Ang Cui, Roberto Di Niro, Jason A Vander Heiden, Adrian W Briggs, Kris Adams, Tamara Gilbert, Kevin C O'Connor, Francois Vigneault, Mark J Shlomchik, and Steven H Kleinstein. A model of somatic hypermutation targeting in mice based on high-throughput ig sequencing data. *The Journal of Immunology*, 197(9):3566–3574, 2016.

[13] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), 2021.

[14] Sivan Eliyahu, Oz Sharabi, Shiri Elmedvi, Reut Timor, Ateret Davidovich, Francois Vigneault, Chris Clouser, Ronen Hope, Assy Nimer, Marius Braun, Yaacov Y. Weiss, Pazit Polak, Gur Yaari, and Meital Gal-Tanamy. Antibody repertoire analysis of hepatitis c virus infections identifies immune signatures associated with spontaneous clearance. *Frontiers in Immunology*, 9:3004, 2018.

[15] Ryan O. Emerson, William S. DeWitt, Marissa Vignali, Jenna Gravley, Joyce K. Hu, Edward J. Osborne, Cindy Desmarais, Mark Klinger, Christopher S. Carlson, John A. Hansen, Mark Rieder, and Harlan S. Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nature Genetics*, 49(5):659–665, May 2017.

[16] Michael Ford, Ehsan Haghshenas, Corey T Watson, and S Cenk Sahinalp. Genotyping and copy number analysis of immunoglobin heavy chain variable genes using long reads. *Iscience*, 23(3):100883, 2020.

[17] Daniel Gadala-Maria, Moriah Gidoni, Susanna Marquez, Jason A Vander Heiden, Justin T Kos, Corey T Watson, Kevin C O'Connor, Gur Yaari, and Steven H Kleinstein. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Frontiers in immunology*, 10:129, 2019.

[18] Daniel Gadala-Maria, Moriah Gidoni, Susanna Marquez, Jason A. Vander Heiden, Justin T. Kos, Corey T. Watson, Kevin C. O'Connor, Gur Yaari, and Steven H. Kleinstein. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Frontiers in Immunology*, 10:129, 2019.

[19] Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, and Steven H Kleinstein. Automated analysis of high-throughput b-cell sequencing data reveals a high frequency of novel immunoglobulin v gene segment alleles. *Proceedings of the National Academy of Sciences*, 112(8):E862–E870, 2015.

[20] Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. Automated analysis of high-throughput b-cell sequencing data reveals a high frequency of novel immunoglobulin v gene segment alleles. *Proceedings of the National Academy of Sciences*, 112(8):E862–E870, 2015.

[21] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, 2014.

[22] Moriah Gidoni, Omri Snir, Ayelet Peres, Pazit Polak, Ida Lindeman, Ivana Mikocziova, Vikas Kumar Sarna, Knut EA Lundin, Christopher Clouser, Francois Vigneault, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by bayesian haplotyping. *Nature communications*, 10(1):1–14, 2019.

[23] Jacob Glanville, Tracy C Kuo, H-Christian von Büdingen, Lin Guey, Jan Berka, Purnima D Sundar, Gabriella Huerta, Gautam R Mehta, Jorge R Oksenberg, Stephen L Hauser, et al. Naive antibody gene-segment frequencies are heritable and unaltered

by chronic lymphocyte ablation. *Proceedings of the National Academy of Sciences*, 108(50):20066–20071, 2011.

[24] Adrian C Hayday, Don J Diamond, Gary Tanigawa, Joseph S Heilig, Virginia Folsom, Haruo Saito, and Susumu Tonegawa. Unusual organization and diversity of t-cell receptor a-chain genes. *Nature*, 316(6031):828–832, 1985.

[25] Stephen M Hedrick, Ellen A Nielsen, Joshua Kavaler, David I Cohen, and Mark M Davis. Sequence relationships between putative t-cell receptor polypeptides and immunoglobulins. *Nature*, 308(5955):153–158, 1984.

[26] Katherine JL Jackson, Bruno A Gaëta, and Andrew M Collins. Identifying highly mutated ighd genes in the junctions of rearranged human immunoglobulin heavy chain genes. *Journal of immunological methods*, 324(1-2):26–37, 2007.

[27] Indu Khatri, Magdalena A. Berkowska, Erik B. van den Akker, Cristina Teodosio, Marcel J.T. Reinders, and Jacques J.M. van Dongen. 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.

[28] Marie J. Kidd, Zhiliang Chen, Yan Wang, Katherine J. Jackson, Lyndon Zhang, Scott D. Boyd, Andrew Z. Fire, Mark M. Tanaka, Bruno A. Gaëta, and Andrew M. Collins. The inference of phased haplotypes for the immunoglobulin h chain v region gene loci by analysis of vdj gene rearrangements. *The Journal of Immunology*, 188(3):1333–1340, 2012.

[29] Ufuk Kirik, Lennart Greiff, Fredrik Levander, and Mats Ohlin. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Molecular immunology*, 87:12–22, 2017.

[30] Marie-Paule Lefranc, Véronique Giudicelli, Chantal Ginestoux, Julia Bodmer, Werner Müller, Ronald Bontrop, Marc Lemaitre, Ansar Malik, Valérie Barbié, and Denys Chaume. Imgt, the international immunogenetics database. *Nucleic Acids Research*, 27(1):209–212, 1999.

[31] Mingfeng Liao, Yang Liu, Jing Yuan, Yanling Wen, Gang Xu, Juanjuan Zhao, Lin Cheng, Jinxiu Li, Xin Wang, Fuxiang Wang, Lei Liu, Ido Amit, Shuye Zhang, and Zheng Zhang. Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nature Medicine*, 26(6):842–844, Jun 2020.

[32] Mao-Jan Lin, Yu-Chun Lin, Nae-Chyun Chen, Allen Chilun Luo, Sheng-Kai Lai, Chia-Lang Hsu, Jacob Shujui Hsu, Chien-Yu Chen, Wei-Shiung Yang, and Pei-Lung Chen. Profiling germline adaptive immune receptor repertoire with gairr suite. *bioRxiv*, 2020.

[33] Shishi Luo, Jane A Yu, Heng Li, and Yun S Song. Worldwide genetic variation of the ighv and trbv immune receptor gene families in humans. *Life Science Alliance*, 2(2), 2019.

[34] Jennifer N Lynch, David L Donermeyer, K Scott Weber, David M Kranz, and Paul M Allen. Subtle changes in tcrα cdr1 profoundly increase the sensitivity of cd4 t cells. *Molecular immunology*, 53(3):283–294, 2013.

[35] Rachel Mackelprang, Robert J Livingston, Michael A Eberle, Christopher S Carlson, Qian Yi, Joshua M Akey, and Deborah A Nickerson. Sequence diversity, natural selection and linkage disequilibrium in the human t cell receptor alpha/delta locus. *Human genetics*, 119(3):255–266, 2006.

[36] Burkhard J Manfras, Dirk Terjung, and Bernhard O Boehm. 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. *Human Immunology*, 60(11):1090 – 1100, 1999.

[37] Vasiliki Matzaraki, Vinod Kumar, Cisca Wijmenga, and Alexandra Zhernakova. The mhc locus and genetic susceptibility to autoimmune and infectious diseases. *Genome biology*, 18(1):1–21, 2017.

[38] Ivana Mikocziova, Moriah Gidoni, Ida Lindeman, Ayelet Peres, Omri Snir, Gur Yaari, and Ludvig M Sollid. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Research*, 48(10):5499–5510, 05 2020.

[39] Ivana Mikocziova, Ayelet Peres, Moriah Gidoni, Victor Greiff, Gur Yaari, and Ludvig M. Sollid. Alternative splice variants and germline polymorphisms in human immunoglobulin light chain genes. *bioRxiv*, 2021.

[40] Kenneth Murphy. *Janeway's Immunobiology*. Garland Science, 8 edition, July 2011.

[41] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2017.

[42] Aviv Omer, Or Shemesh, Ayelet Peres, Pazit Polak, Adrian J Shepherd, Corey T Watson, Scott D Boyd, Andrew M Collins, William Lees, and Gur Yaari. Vdjbase: an adaptive immune receptor genotype and haplotype database. *Nucleic acids research*, 48(D1):D1051–D1056, 2020.

[43] R-Y Pan, R-L Dao, S-I Hung, and W-H Chung. Pharmacogenomic advances in the prediction and prevention of cutaneous idiosyncratic drug reactions. *Clinical Pharmacology & Therapeutics*, 102(1):86–97, 2017.

[44] Ren-You Pan, Mu-Tzu Chu, Chuang-Wei Wang, Yun-Shien Lee, Francois Lemonnier, Aaron W Michels, Ryan Schutte, David A Ostrov, Chun-Bing Chen, Elizabeth Jane Phillips, et al. Identification of drug-specific public tcr driving severe cutaneous adverse reactions. *Nature communications*, 10(1):1–13, 2019.

[45] Ayelet Peres, Moriah Gidoni, Pazit Polak, and Gur Yaari. RAbHIT: R Antibody Haplotype Inference Tool. *Bioinformatics*, 06 2019.

[46] Duncan K Ralph and Frederick A Matsen IV. Consistency of vdj rearrangement and substitution parameters enables accurate b cell receptor sequence annotation. *PLoS computational biology*, 12(1):e1004409, 2016.

[47] Duncan K Ralph and Frederick A Matsen IV. Per-sample immunoglobulin germline inference from b cell receptor deep sequencing data. *PLoS computational biology*, 15(7):e1007133, 2019.

[48] Harlan S. Robins, Paulo V. Campregher, Santosh K. Srivastava, Abigail Wacher, Cameron J. Turtle, Orsalem Kahsai, Stanley R. Riddell, Edus H. Warren, and Christopher S. Carlson. Comprehensive assessment of t-cell receptor beta-chain diversity in alphabeta t cells. *Blood*, 114(19):4099–4107, Nov 2009. PMC2774550[pmcid].

[49] Oscar L. Rodriguez, William S. Gibson, Tom Parks, Matthew Emery, James Powell, Maya Strahl, Gintaras Deikus, Kathryn Auckland, Evan E. Eichler, Wayne A. Marasco, Robert Sebra, Andrew J. Sharp, Melissa L. Smith, Ali Bashir, and Corey T. Watson. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Frontiers in Immunology*, 11:2136, 2020.

[50] Elisa Rosati, C Marie Dowds, Evaggelia Liaskou, Eva Kristine Klemsdal Henriksen, Tom H Karlsen, and Andre Franke. Overview of methodologies for t-cell receptor repertoire analysis. *BMC biotechnology*, 17(1):1–16, 2017.

[51] Lee Rowen, Ben F. Koop, and Leroy Hood. The complete 685-kilobase dna sequence of the humanβ t cell receptor locus. *Science*, 272(5269):1755–1762, 1996.

[52] Florian Rubelt, Christopher R Bolen, Helen M McGuire, Jason A Vander Heiden, Daniel Gadala-Maria, Mikhail Levin, Ghia M Euskirchen, Murad R Mamedov, Gary E Swan, Cornelia L Dekker, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nature communications*, 7(1):1–12, 2016.

[53] Daniel B Rubinstein, Michel Symann, A Keith Stewart, and Thierry Guillaume. Restriction fragment length polymorphisms and single germline coding region sequence in vh182, a duplicated gene encoding autoantibody. *Molecular immunology*, 30(4):403–412, 1993.

[54] EH Sasso, JH Buckner, and LA Suzuki. Ethnic differences in vh gene polymorphism. *Annals of the New York Academy of Sciences*, 764(1):72–73, 1995.

[55] Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vince Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, Will Chow, Joanna Collins, Glenn Harden, Tim Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, and Deanna M. Church. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *bioRxiv*, 2016.

[56] EK Shin, F Matsuda, H Nagaoka, Y Fukita, T Imai, K Yokoyama, E Soeda, and T Honjo. Physical map of the 3' region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. *The EMBO journal*, 10(12):3641–3645, 1991.

[57] Euy Kyun Shin, Fumihiko Matsuda, Shoichi Ozaki, Shun-ichi Kumagai, Olle Olerup, Håkan Ström, Inga Melchers, and Tasuku Honjo. Polymorphism of the human immunoglobulin variable region segment v1-4.1. *Immunogenetics*, 38(4):304–306, 1993.

[58] Donjete Simnica, Nuray Akyüz, Simon Schliffke, Malte Mohme, Lisa v.Wenserski, Thorben Mährle, Lorenzo F. Fanchi, Katrin Lamszus, and Mascha Binder. T cell receptor next-generation sequencing reveals cancer-associated repertoire metrics and reconstitution after chemotherapy in patients with hematological and solid tumors. *OncoImmunology*, 8(11):e1644110, 2019. PMID: 31646093.

[59] Natanael Spisak, Aleksandra M Walczak, and Thierry Mora. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic acids research*, 48(19):10702–10712, 2020.

[60] Lakshman Subrahmanyan, Michael A Eberle, Andrew G Clark, Leonid Kruglyak, and Deborah A Nickerson. Sequence variation and linkage disequilibrium in the human t-cell receptor β (tcrb) locus. *The American Journal of Human Genetics*, 69(2):381–395, 2001.

[61] Barry Toyonaga, Yasunobu Yoshikai, Veronica Vadasz, Beth Chin, and Tak W Mak. Organization and sequences of the diversity, joining, and constant region genes of the human t-cell receptor beta chain. *Proceedings of the National Academy of Sciences*, 82(24):8624–8628, 1985.

[62] J. J. M. van Dongen, A. W. Langerak, M. Brüggemann, P. A. S. Evans, M. Hummel, F. L. Lavender, E. Delabesse, F. Davi, E. Schuuring, R. García-Sanz, J. H. J. M. van Krieken, J. Droese, D. González, C. Bastard, H. E. White, M. Spaargaren, M. González, A. Parreira, J. L. Smith, G. J. Morgan, M. Kneba, and E. A. Macintyre. 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*, 17(12):2257–2317, Dec 2003.

[63] Jason A. Vander Heiden, Gur Yaari, Mohamed Uduman, Joel N.H. Stern, Kevin C. O'Connor, David A. Hafler, Francois Vigneault, and Steven H. Kleinstein. presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, 30(13):1930–1932, 2014.

[64] Morgan E Wallace, Michelle Bryden, Stephen C Cose, Richard M Coles, Ton N Schumacher, Andrew Brooks, and Francis R Carbone. Junctional biases in the naive TCR repertoire control the CTL response to an immunodominant determinant of HSV-1. *Immunity*, 12(5):547–556, May 2000.

[65] Corey T Watson, Frederick A Matsen 4th, Katherine JL Jackson, Ali Bashir, Melissa Laird Smith, Jacob Glanville, Felix Breden, Steven H Kleinstein, Andrew M Collins, and Christian E Busse. Comment on" a database of human immune receptor alleles recovered from population sequencing data". *Journal of immunology (Baltimore, Md.: 1950)*, 198(9):3371–3373, 2017.

[66] Corey T Watson, Karyn M Steinberg, John Huddleston, Rene L Warren, Maika Malig, Jacqueline Schein, A Jeremy Willsey, Jeffrey B Joy, Jamie K Scott, Tina A Graves, Richard K Wilson, Robert A Holt, Evan E Eichler, and Felix Breden. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *The American Journal of Human Genetics*, 92(4):530–546, 2013.

[67] Wen Wen, Wenru Su, Hao Tang, Wenqing Le, Xiaopeng Zhang, Yingfeng Zheng, Xiuxing Liu, Lihui Xie, Jianmin Li, Jinguo Ye, Liwei Dong, Xiuliang Cui, Yushan Miao, Depeng Wang, Jiantao Dong, Chuanle Xiao, Wei Chen, and Hongyang Wang. Immune cell profiling of covid-19 patients in the recovery stageby single-cell sequencing. *Cell Discovery*, 6(1):31, May 2020.

[68] Gur Yaari, Jason Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel Stern, Kevin O'Connor, David Hafler, Uri Laserson, Francois Vigneault, and

Steven Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology*, 4:358, 2013.

[69] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. Igblast: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40, 2013.

[70] Yasunobu Yoshikai, Stephen P Clark, Sheryle Taylor, Uik Sohn, Bonnie I Wilson, Mark D Minden, and Tak W Mak. Organization and sequences of the variable, joining and constant region genes of the human t-cell receptor α-chain. *Nature*, 316(6031):837–840, 1985.

[71] Yaxuan Yu, Rhodri Ceredig, and Cathal Seoighe. Lymanalyzer: a tool for comprehensive analysis of next generation sequencing data of t cell receptors and immunoglobulins. *Nucleic acids research*, 44(4):e31–e31, 2016.

[72] Yaxuan Yu, Rhodri Ceredig, and Cathal Seoighe. A database of human immune receptor alleles recovered from population sequencing data. *The Journal of Immunology*, 198(5):2202–2210, 2017.

[73] T M Zhao, S E Whitaker, and M A Robinson. A genetically determined insertion/deletion related polymorphism in human T cell receptor beta chain (TCRB) includes functional variable gene segments. *Journal of Experimental Medicine*, 180(4):1405–1414, 10 1994.