pmTR database: population matched (PM) germline allelic variants of T-cell receptor (TR) loci

Julian Dekker
Leiden University Medical Center
Jacques J.M. Dongen (J.J.M.van_Dongen@lumc.nl)
Leiden University Medical Center
Marcel J.T. Reinders
Delft University of Technology
Indu Khatri
Leiden University Medical Center

Research Article

Keywords: population, germline, allelic variants, TR loci, diversity

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-234725/v1

Version of Record: A version of this preprint was published at Genes & Immunity on April 18th, 2022.
See the published version at https://doi.org/10.1038/s41435-022-00171-x.

Abstract

T-cell receptor (*TR*) germline allele sequences are arranged, organized and made available to the research community by the IMGT database. This state-of-the-art database, however, does not provide information regarding population specificity and allelic frequencies of the four human *TR* loci (*TRA*, *TRB*, *TRG* and *TRD*). The specificity of allelic variants to different human populations can, however, be a rich source of information when studying the genetic basis of population-specific immune responses in disease and in vaccination. To make *TR* germline alleles available for such population-specific studies, we meticulously identified true germline alleles enriched with complete *TR* allele sequences and their frequencies across 26 different human populations, profiled by “1,000 Genomes data”. We identified 205 *TRAV*, 249 *TRBV*, 16 *TRGV* and 5 *TRDV* germline alleles supported by at least four haplotypes (= minimum of two unrelated individuals). The diversity of germline allelic variants in the *TR* loci is highest in Africans, while the majority of the Non-African alleles are specific to the Asian populations, suggesting a diverse profile of *TR* germline alleles in different human populations. Interestingly, the alleles known in the IMGT database are frequent and common across all five super-populations. We believe that this new set of genuine germline *TR* sequences represents a valuable new resource which we have made available through the new population-matched *TR* (pmTR) database, accessible via https://pmtrig.lumc.nl/.

Introduction

The genomic organization of four loci of T-cell receptors (TR), i.e. alpha (*TRA*), beta (*TRB*), gamma (*TRG*) and delta (*TRD*), is complex. The four loci are distributed over three different genomic regions across two chromosomes in the human genome: *TRA* and *TRD* are located intermingled on chromosome 14q11.2 (with *TRD* embedded within the *TRA* locus), and *TRB* and *TRG* on different arms of chromosome 7. The *TRB* and *TRD* loci are comprised of *Variable* (*V*), *Diversity* (*D*) and *Joining* (*J*) genes, whereas *TRA* and *TRG* loci contain *V* and *J* genes only. The polypeptides, encoded by functionally rearranged *TRA* and *TRB* loci, combine to form a TRαβ receptor, whereas functionally rearranged *TRD* and *TRG* loci form the TRγδ receptor, both containing an antigen-recognition domain. The recombination of *V(D)J* genes has the potential to generate many millions of different TR molecules, each having a unique antigen binding specificity [1]. The *V(D)J* recombination process is directed by “recombination signal sequences” (RSS), short highly conserved DNA stretches, present at each recombination site of the *TR* genes, i.e. downstream to *V*, upstream to *J*, and at both sites of *D* [2,3].

TR genes harbor inter-individual germline allelic variants, causing different individuals to be able to produce different receptors. As these different allelic variants are shared within confined human populations [4], they contribute also to more extreme diversity of receptors at the population level [5]. These population-specific germline variations have been shown to introduce varying disease prevalences in specific population [6–9]. For example, in Asian and Caucasian populations, *TRBV17* plays a pivotal role in Influenza A virus specific T-cell immunity [10]. Consequently, to understand (population-specific) immune responses, a catalogue of population-wide observed TR alleles is crucial. Till today, there is, however, only

one database that reports all alleles for the *TR* loci: the International ImMunoGeneTics information system (IMGT) [11,12]. But, this database does not report allelic frequencies or population statistics and, moreover, reported alleles are mostly profiled from Caucasian populations [13,14].

To enrich the catalogue of TR germline genes with population information, we relied on the "1,000 Genomes (G1K)" dataset (https://www.internationalgenome.org/), derived from cell samples of 2,548 individuals across five different ethnicities. We are not the first in doing so. Yu *et al* created the Lym1k database for immunoglobulin (*IG)* and *TR* loci, also from the G1K data using their AlleleMiner tool [14]. They, however, did not provide any information on the reliability of the (newly) identified alleles and also the link to population information was not retained. Moreover, not all relevant components of each *TR* locus were stored, i.e. they neglected the *D*, *J*, *C* genes and the RSSs. Also, they were not able to profile all *TR* genes as they used a previous version of the G1K dataset (i.e. a mapping to GRCh37 being liftover to GRCh38).

Here, we identified the alleles for all components of all four *TR* loci, i.e. the *V*, *D*, *J*, *C* genes as well as the RSSs, report reliability scores for the differently detected alleles as well as population information of each allele, and present an online accessible database containing this information which we called the "population-matched germline allelic variants of T-cell receptor loci" database; or in short, the *pmTR* database. To realize this, we have developed an automated pipeline to profile all the *TR* alleles from the G1K data. The pipeline returns the sequences of alleles, frequency of alleles, as well as the population distribution of each allele among 26 different populations profiled in G1K resource. The resulting alleles are manually curated and made available via GitHub and the online database (www.pmTRIG.com), including population information and confidence levels to provide access for the community. We have also enabled a BLAST search on the database to directly use our germline alleles in further research.

Results

Population matched germline TR **alleles (pmTR) database application**

We identified population-specific alleles in all four *TR* loci (*TRA*, *TRB, TRG* and *TRD*) using pmAlleleFinder pipeline (Methods), where each allele is supported by at least four haplotypes for 2,548 individuals belonging to 26 populations representing five continents (**Table S1**). All the alleles identified from G1K are identified to create a population matched *TR* (pmTR) database. We identified two to three times more new alleles then present in the standard reference IMGT database for all the genes in the variable genes of all *TRA* and *TRB* loci (Table 1). These alleles were divided into three allele sets (AS1, AS2, or AS3) based on different confidence levels (Methods). AS1 are alleles already present in the IMGT database; AS2 are novel alleles that are frequent in the populations (supported by at least 19 haplotypes, i.e. 10 or more individuals); and AS3 are novel but rare alleles (supported by 4 to 18 haplotypes, i.e. at least 2 to 9 individuals). The pmTR database further contains meta information about the alleles such as the support of haplotypes for each (sub)population (**Tables S2, S3, S4, and S5)**. We have also identified the Recombination signal sequences (RSS) for each gene and the corresponding variants for each allele. The

heptamers and nonamers in the RSS sequence turned out not to be conserved for most *TR* genes resulting in lower recombinant frequencies for genes with less conserved RSS [15]. Similar to *IG* pseudogenes [13], we also identified conserved heptamers for *TR* pseudogenes, suggesting the role of RSS in recombination of pseudogenes in T-cell repertoire [16] (**Table S6**).

Table 1
Number of alleles in different functional gene segments in IG **loci.** AS1 (Known), AS2 (Frequent) and AS3 (Rare) are major confidence levels.

	AS1	AS2	AS3	Total	IMGT
TRAV	54	75	76	205	87
TRAJ	64	18	17	99	71
TRAC	1	2	1	4	1
TRBV	68	82	99	249	118
TRBD	1	0	0	1	3
TRBJ	16	1	0	17	16
TRBC	1	2	3	6	4
TRGV	13	30	23	66	16
TRGJ	5	2	1	8	6
TRGC	13	8	7	28	13
TRDV	5	5	2	12	5
TRDD	3	0	0	3	3
TRDJ	4	2	0	6	4
TRDC	1	0	2	3	1

Variable alleles in the IMGT database are partial and are frequently present in all ethnicities

Mapping the pmTR alleles to the known alleles in the IMGT database is instrumental in this setting for assessing the frequency and the population specificity of the known alleles, as such information will be helpful in understanding the population-specific response in disease and to vaccines. We found that 60–100% of the IMGT alleles for functional variable genes in of the each *TR* loci were identified in the pmTR database. The functional *V* gene alleles in the IMGT database are not complete as we found 42 of 87 *TRAV* and 53 of 118 *TRBV* alleles to be partial i.e. they do not comprise a leader sequence. In most cases only the first, or first two, alleles of the *TR* genes are sequenced completely. Moreover, the majority of the

TR genes in the IMGT had only one allele recorded, implying that a complete *TR* germline allele resource is not available for the research community. When mapping pmTR and IMGT alleles only over the *V* exon region, an increase in the known alleles (AS1 category) was observed. Moreover, looking at the super-population distribution of the IMGT alleles in our pmTR database, we found that a majority of these alleles (> 90%) are shared among all the super-populations (Fig. 1). Moreover, more than 90% of the mapped IMGT alleles are supported with at least 100 haplotypes and are frequently present in all the super-populations (Fig. 1, **Table S2-S5**).

The majority of the novel V gene alleles are one mutation away from known alleles

To gain information on the number of new mutations added by the novel *V* gene alleles to the mutating positions in the known germline alleles, it was important to estimate the differences that are added by the mutations in the Leader region and *V* exon separately as only *V* exon rearranges with *(D)J* genes to generate T-cell repertoire. For complete *V* genes (Leader + *V* exon), we found that 81% (113/139) of the novel *TRAV* alleles and 88% (142/161) of the novel *TRBV* alleles are one mutation away from their known alleles (Table 2), i.e. they only have one different mutation with respect to a known IMGT allele. These mutated positions are randomly distributed over the *V* gene and do not show specific patterns.

Table 2
Number of alleles with count of new mutating positions as compared to the existing databases.
Complete sequence includes leader sequence and the *V* region for all the *V* genes/ This is important to realize that 42 *TRAV* and 53 *TRBV* alleles are partial and do not contain the leader sequence.

Mutating positions	TRAV		TRBV	
	Complete sequence	V region only	Complete sequence	V region only
0	66	71	88	104
1	113	114	142	132
2	22	17	14	10
3	3	2	5	3
4	1	1	0	0

Novel alleles are unique to super-populations

Opposite to the known alleles (AS1 category), which are shared between all the population, novel alleles for *TR* loci (AS2 and AS3 categories) are unique to specific super-populations. Remarkably, very few novel *TR* alleles are shared among all five super-populations. One-third of the novel *TRAV* alleles constitute of African-specific alleles (Fig. 2A). Two-third of these novel African specific alleles are rare, i.e. supported by less than 19 haplotypes. In line with this, 32 of the 34 population-specific *TRAJ* alleles are not known in the IMGT database. Furthermore, we found an uneven diversity in *TRAV* genes, e.g. *TRAV40*, the only known allele of 11 alleles is rare in human populations and we found that each new allele is unique to a

particular population (Fig. 2B). Several novel alleles for gene *TRAV8-4*, *V38-1*, *V27* are unique to African populations. Moreover, we also observed frequent novel alleles for *TRAV30* and *TRAV8-2* genes.

TRBV follows a similar pattern as the *TRAV* alleles; half of the novel alleles are African-specific of which a majority is rarely present in Africans (Fig. 3A). However, we do not observe a similar pattern for the *TRBD*, *TRBJ* and *TRBC* alleles. Similar to *TRAV* population distribution, we observed a specific pattern in diversity of *TRBV* genes. The novel alleles in *TRBV5-4* and *V7-9* genes consists several alleles unique to African populations whereas *TRBV5-6*, *V15* and *V23* genes comprises alleles shared between African and American populations (Fig. 3B). On the contrary, *TRBV6-7* gene comprises novel South Asian alleles. In general, half of the 'AFR shared' *TRAV* and *TRBV* alleles are common between African and American super-populations (Fig. 2B and 3B), suggesting a role of intermixing due to migratory events.

TRG and *TRD*, being the smallest *TR* loci, have fewer novel alleles as compared to the *TRA* and *TRB* loci (Fig. 4A). However, unlike *TRAV* and *TRBV* novel alleles, novel *TRGV* alleles often belong to Non-African populations (Fig. 4A). We found a higher diversity in genes *TRGV3*, *V5* and *V9* as compared to other genes suggesting the role of evolutionary pressure on these genes (Fig. 4B). Moreover, the *TRG* locus has the highest number of alleles for constant genes across all four *TR* loci, a majority of which belongs to the rare category (AS3) (Fig. 4A). The *TRD* locus is the most conserved locus as very few novel alleles were found for the *TRDV*, *TRDJ* and *TRDC* genes (Fig. 4A). An equal distribution of African specific and non-African specific alleles was observed in *TRDV* alleles, wherein the majority of non-African alleles were specific to European and South Asian populations (Fig. 4C). Interestingly, similar to *TRGV*, we also found a more diverse pattern in *TRDV2* gene as compared to the *TRDV1* and *TRDV3*.

Summarizing, we find a similar super-population distribution across all four *TR* loci despite their size difference, peculiar gene-specific population distribution and varying levels of duplication.

A majority of Non-African alleles are specific to Asian super-populations

The super-population distribution of novel *TRA*, *TRB* and *TRG* alleles show that these are specific to the Non-African populations (Figs. 2–4). In fact, a majority of them belongs to the East and South Asian super-populations (Fig. 5). Interestingly, these alleles are not specific to any of the Asian populations. Very few alleles were shared between Asian and American-European super-populations, suggesting an exclusive nature of the *TR* loci in these ethnicities. The larger number of Asian-specific alleles suggests an exceptional diversity in Asia which went unnoticed in the current databases.

Genomic organization of the TR loci governs the evolutionary dynamics

A principle component analysis (PCA) was performed on the entire span of all four *TR* loci to visualize the genetic diversity among different loci. We found that the Africans are highly diverse in all the four loci, whereas other super-populations are comparatively similar to each other (Fig. 6). Interestingly *TRA* and *TRB* follow similar patterns despite being on different chromosomes. On the other hand, in the *TRG* and

TRD loci, we observed multiple clusters of individuals from all the super-populations having unique variability as compared to the variation in the individuals belonging to African populations. Despite that the *TRD* locus lies within the *TRA* locus, it does not follow a similar variability as the *TRA* locus, implying that selection pressure has been different when comparing the *TRA* and *TRB* loci with the *TRG* and *TRD* loci. This may have been governed by the size of loci and particularly by different functional aspects of the TCRαβ vs. TCRγδ molecules.

To investigate the population structure in more detail, we calculated the population differentiation for each *TR* loci separately. We found a comparatively different structure between the loci as compared to the genetic diversity assessed using PCA in Fig. 6A. We found Africans to be the most diverse in the *TRB* and *TRG* loci whereas Americans and Europeans are ancestor clades for the *TRA* and *TRD* loci. The population structure is in accordance with the chromosomal organization of these loci (Fig. 6B), unlike the genetic diversity visualized in the PCA plot. Interestingly, in all the *TR* loci, an early separation of Mexicans (MXL) and Peruvians (PEL) populations is observed (Fig. 6B). Similar to the population structure in *TRG* loci, the relatedness of the PEL population was also observed in the *IG* loci [13].

Discussion

We performed a comprehensive analysis of *TR* germline alleles identified from 2,548 individuals available in the 1,000 Genomes data belonging to 26 populations across five different ethnicities. The *TR* germline alleles from the 1,000 Genomes data were also identified previously by Yu et al. and compiled in the Lym1K resource. This resource, however, does not provide information on the allelic frequencies and population-specificities of the alleles [14]. In addition, we not only used a minimum haplotype count (i.e. 4 haplotypes) to identify alleles using our automated pipeline “pmAlleleFinder”, but also categorized the pmTR alleles in three confidence levels. Moreover, potential false-positive germline alleles in operationally indistinguishable (OI) genes were also eliminated by manually assessing the mutating positions between the alleles belonging to OI genes [13]. Finally, the filtered alleles for *TR* loci were compiled in the pmTR database (hosted via https://pmtrig.lumc.nl/). Application of these rules and checks, has provided us with genuine *TR* germline alleles.

The assignment of the alleles to three confidence categories provided us with a notion of accuracy for the pmTR alleles. 60–100% of the IMGT alleles for *Variable* genes were also recorded in the pmTR database. The coverage even increased to > 80% of the IMGT alleles when the mapping was performed across the *V* exon region only. A complete overlap between the pmTR and IMGT *Variable* gene alleles is hampered by the number of individuals with which the pmTR database is created, i.e. 2,548, as compared to thousands of individuals profiled by the IMGT database over last decade. However, in the pmTR database, all genes/alleles from one individual are profiled, whereas the IMGT database has been compiled with the few (rearranged) genes/alleles of many different individuals. We found that > 90% of the IMGT alleles are shared between all five super-populations, with very few alleles being specific to one or more of the super-populations, suggesting that the IMGT database lacks rare population-specific alleles. Having only alleles that are shared uniformly and frequently across all super-populations, poses limitations for immune-

response studies that aim at understanding the genetic basis of difference between and among different human populations [4,17].

Most of the novel pmTR alleles were identified in African populations. >90% of these novel alleles are not captured by the IMGT database. Moreover, we found several rare African alleles in all the four *TR* loci. This huge diversity can be explained by the recent genomics-studies in the African populations [18–20]. The sampling of several individuals from many African populations can substantiate our understanding of allelic diversity in *TR* loci. In contrast to *IG* alleles [13], where African-Shared alleles were the second largest group of alleles [13], the Non-African alleles are the second largest group for *TR* alleles. A majority of these Non-African alleles are specific to the Asian populations, suggesting a divergence of *TR* loci, resulting in unique alleles across different ethnicities. We also observed a similar trend in the evolutionary dynamics of the independent *TR* loci represented by the genetic diversity in the PCA plots. The African individuals showed the highest genetic diversity in all the loci. However, we found that the genomic organization governs the population structures of the four *TR* loci. *TRA* and *TRD* showed similar patterns with the Americans and Europeans being the most diverse, whereas the African populations are the most variable for the *TRB* and *TRG* loci. Here it should be noted that the *TRD* and *TRG* loci are the smallest loci, and, hence, the genetic diversity and the population structure can be affected by the size of the loci.

Gleaning into the gene specific population distribution of the *TR* alleles, we found a unique pattern in several genes. Some genes had a higher diversity in alleles in preferred populations (e.g. alleles for *TRAV8-4*, TRAV27, *TRBV5-4*, *TRBV7-9* genes comprise African specific alleles) whereas some tend to be conserved across all the populations (e.g. alleles for *TRAV18, TRAV8-6, TRBV24-1, TRGV1, TRDV1* genes). *Diversity*, *Joining* and *Constant* genes in general were conserved suggesting lower level of evolutionary selection pressure on these genes. Moreover, some genes tend to have a huge diversity as compared to other genes (e.g. *TRAV40*, *TRBV7-9*, *TRGV3, TRGV5, TRDV1*). The *TRGV9* and *TRDV2* are heavily expanded in the first few years of life [21]. The high diversity of these genes at population level suggest a specific selection process at the population level as well. All these interesting findings suggest the role of underlying selection mechanisms possibly owing to migratory events of the human populations [22,23]. The unique diversity of these genes can result in the preferential usage of specific alleles in selective populations ultimately shaping the population-specific immune responses.

Taken together, we meticulously reported curated germline alleles across 5 ethnicities containing 26 subpopulations, resulting in ~ 150% more alleles as compared to the known alleles within the IMGT database. This enriched resource can be used for the repertoire studies to understand (population-specific) immune response dynamics.

Methods

Data source

The 1,000 Genomes data (G1K) (March, 2019 release; http://www.1000genomes.org; GRCh38 assembly) in the form of phased variant cell format (VCF) was used for retrieving the *TR* germline alleles. Phased variants for GRCh38 (a recent release for the 1,000 genomes) were used as they were the most recent and comprised almost all the genes of the *TR* locus as compared to the GRCh37 assembly. The full release of the data set was collected from 2,548 cell samples from diverse ethnic groups that have a uniform distribution of individuals across populations. The samples are classified in five super populations i.e. Africa, America, East Asia, Europe and South Asia, that are further subdivided into 26 populations (7 African, 4 American, 5 East Asian, 5 European and 5 South Asian populations) with a minimum of 61 and a maximum of 113 samples per population (**Table S1**). The phased VCF format of the data comprises information of both parental (forward) and maternal (reverse) chromosomes for each sample.

Terminology and Nomenclature

With the term *"haplotype*" we refer to a gene present on one strand (inherited from a single parent) in one individual. Therefore, there are two haplotypes, one on the positive and one on the negative strand, with exactly the same or different polymorphisms. *"Allele"* refers to a haplotype from multiple individuals consisting of the same variants across the complete gene sequence. *"Mutations"* are genetic mutations that occurred to form different alleles (also denoted as allelic variants) of the same gene. The IMGT nomenclature is used to name genes, and this name is extended with a numbering to refer to the different alleles, for example the 01 and 02 alleles of the *TRAV1* gene are referred to as *TRAV1_01*, *TRAV1_02*. IMGT alleles are denoted with an asterix, such as *TRAV1*01*, *TRAV1*02*. The alleles were sorted in descending order such that the first allele is supported by the maximum number of haplotypes.

An automated pipeline to identify germline alleles from G1K data

The pmAlleleFinder pipeline is used to infer the alleles of the genes of interest i.e. *V*, *D*, *J* and *C* genes and RSSs from the input VCF file. The pipeline results in a list of alleles for each gene separately along with the population information of each allele in a separate file (**Figure S1**). It finds all possible haplotypes for each gene, merges them into alleles and counts the haplotype frequency of each such allele. The alleles can be filtered by the user through defining a minimum number of haplotype support. The pipeline is developed in python and R with an additional possibility to automatically identify if pmTR alleles are present in the IMGT database. The pipeline is not limited to the identification of alleles for *TR* genes only and, given a phased vcf file, can be used to find population-based alleles for any gene.

Allele confidence levels

The alleles obtained from the G1K resource were classified into three major confidence levels (allele set (AS) 1–3):

AS1 (known)

G1K alleles with a minimum support of 4 haplotypes and identified in the IMGT databases. This AS1 allele set has the highest level of confidence as the alleles are observed in the G1K resource as well as in the IMGT database.

AS2 (frequent novel alleles)

G1K alleles with a minimum support of 19 haplotype (at least ten individuals). These alleles represent a set of newly identified alleles that are frequent.

AS3 (rare novel alleles)

G1K alleles that have a haplotype support between 4 and 18 (supported by two to nine individuals). Despite the rarity of these alleles, we believe them to be genuine as the chance that 4 *identical* haplotypes within 5,096 independent haplotypes are caused by sequencing errors is highly unlikely.

As few *V* genes are paralogous [17,24], the mapping of short reads to such genes can be erroneous, influencing the subsequently derived alleles. Called mutations on the alleles of such genes can thus easily be false positives, even after using stringent parameters. Therefore, we denote these genes as *operationally indistinguishable* (OI) genes. As these genes can be recognized based on their sequence similarity [25], we generated a neighbor-joining (NJ) tree for all *V* genes on the *TRA*, *TRB, TRG* and *TRD* loci, separately. The genes sharing a clade with a short branch length, i.e. 0.02, are called OI genes (**Figure S2**); and the corresponding alleles OI alleles.

Filtering false positive alleles

The G1K alleles were scrutinized manually: 1) alleles with stop codons were removed from the final set, and 2) alleles within OI genes where removed when they had a mutation that is shared with alleles for the other OI genes as it points towards a mis-alignment of a read (when the mutation is present in the IMGT databases across multiple alleles it is not filtered).

Population annotation of alleles

G1K alleles are annotated with super-population information (**Tables S2-S5**) into four categories: 1) ALL, present in all super-populations; 2) AFR, only present in Africans; 3) AFR SHARED, present in African and at least one of the other super-populations, but not all; and 4) NON-AFR, present in at least one of the super-populations but not in Africans.

Phylogenetic trees for alleles

Maximum Likelihood (ML) trees were built for the alleles using RAxML [26]. The PROTGAMMAJTT model was used to build the trees with 100 bootstraps. The trees were visualized using the iTOL server [27]. The trees taxa were colored as per AS classification; the population level annotation is displayed in binary format and the frequency of alleles as text. A few alleles derived from loci not meant for evaluation were used as an outgroup in all the ML trees.

pmTR online database

The online pmTR database front-end is made with ReactJS in combination with the Neo4j graph database back-end to load and display all genes [28,29], respective alleles, population frequencies and confidence levels (AS1-AS3). Genes can be searched on the basis of their name or nucleotide sequence enabled by a BLAST search [30] (**Figure S3**). The pmTR is hosted on https://pmtrig.lumc.nl/.

Genetic diversity and migration events

The VCF file of the complete individual locus, i.e. *TRA* (Chr14 [21621904, 22552132), *TRB* (Chr7 [142299011, 142813287]), *TRG* (Chr7 [38240024, 38368055]) and *TRD* (Chr14 [22422546, 22466577]) was obtained. The SNPs from the coding and non-coding region from each locus were independently subjected to a principal component analysis (PCA) using the R Bioconductor package 'SNPRelate' [31]. The pairwise population differentiation, quantified by the fixation index (F_{ST}), was calculated based on levels of differentiation in polymorphism frequencies across populations. F_{ST} is proportional to the evolutionary branch length between each pair of populations. A Neighbor joining tree was used to visualize the F_{ST} distances between populations.

Declarations

Acknowledgements

We acknowledge 1,000 Genomes project for making the data publicly available. As we have used a specific locus of the chromosomes 7 and 14, we have complied with the 1,000 Genomes policies for the publication of data.

Authors contributions statement

The study was conceptualized by JJMvD. The pipeline was developed by IK which was automated by JD. JD performed the data acquisition, data analysis and organization and development of the database. IK and MJTR supervised the analysis. All the authors wrote the manuscript and designed the figures. All the authors approved the final version of the manuscript.

Competing interests

JJMvD is the founder of the EuroClonality Consortium and one of the inventors on the EuroClonality-owned patents and EuroFlow-owned patents, which are licensed to Invivoscribe, BD Biosciences or Cytognos; these companies pay royalties to the EuroClonality and EuroFlow Consortia, respectively, which are exclusively used for sustainability of these consortia. JJMvD reports an Educational Services Agreement with BD Biosciences and a Scientific Advisory Agreement with Cytognos to LUMC.

The rest of the authors declare that they have no other relevant conflicts of interest.

Funding disclosure

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 707404. The opinions expressed in this document reflect only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

This project has also received funding from the PERISCOPE program. PERISCOPE has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115910. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industries and Associations (EFPIA) and Bill and Melinda Gates Foundation (BMGF).

Availability of data

The source code of the automated tool to identify population-matched alleles and automated mapping to known resources is available via GitHub (https://github.com/JulianDekker/PMalleleFinder). The database is hosted via a website at (www.pmTRIG.com).

References

1. Parham, P. *The immune system*. (Garland Science, 2015).
2. Bassing, C. H. *et al.* Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule. *Nature***405**, 583–586 (2000).
3. Mcblane, J. F. *et al.Cleavage at a V(D)J Recombination Signal Requires Only RAG1 and RAG2 Proteins and Occurs in Two Steps*. *Cell***83**, (1995).
4. De Inocencio, J., Choi, E., Glass, D. N. & Hirsch, R. T cell receptor repertoire differences between African Americans and Caucasians associated with polymorphism of the TCRBV3S1 (V beta 3.1) gene. *J. Immunol.***154**, 4836–41 (1995).
5. Rosenberg, W. M. C., Moss, P. A. H. & Bell, J. I. Variation in human T cell receptor Vβ and Jβ repertoire: analysis using anchor polymerase chain reaction. *Eur. J. Immunol.***22**, 541–549 (1992).
6. Chang, W. C. *et al.* V-J combinations of T-cell receptor predict responses to erythropoietin in end-stage renal disease patients. *J. Biomed. Sci.***24**, 43 (2017).
7. Peng, W. *et al.* Profiling the TRB and IGH repertoire of patients with H5N6 Avian Influenza Virus Infection by high-throughput sequencing. *Sci. Rep.***9**, 1–11 (2019).
8. Shi, B. *et al.* Compositional characteristics of human peripheral TRBV pseudogene rearrangements. *Sci. Rep.***8**, (2018).
9. Cui, H. *et al.* Analysis of differential β variable region of T cell receptor expression and NAV3/TNFRSF1B gene mutation in mycosis fungoides. *Oncotarget***7**, 17986–17990 (2016).

10. Liu, J. *et al.* Conserved epitopes dominate cross-CD8+ T-cell responses against influenza A H1N1 virus among Asian populations. *Eur. J. Immunol.***43**, 2055–2069 (2013).
11. Lefranc, M.-P. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.***29**, 207–209 (2001).
12. Lefranc, M.-P. *et al.* IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.***27**, 209–212 (1999).
13. Khatri, I. *et al.* 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* 2020.04.09.033530 (2020). doi:10.1101/2020.04.09.033530
14. Yu, Y., Ceredig, R. & Seoighe, C. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. *J. Immunol.***198**, 2202–2210 (2017).
15. Hesse, J. E., Lieber, M. R., Mizuuchi, K. & Gellert, M. V(D)J recombination: a functional definition of the joining signals. *Genes Dev.***3**, 1053–61 (1989).
16. Shi, B. *et al.* Compositional characteristics of human peripheral TRBV pseudogene rearrangements. *Sci. Rep.***8**, 1–9 (2018).
17. Watson, C. T. *et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.***92**, 530–46 (2013).
18. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0273-y
19. Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell***167**, 657-669.e21 (2016).
20. Fatumo, S. The opportunity in African genome resource for precision medicine. *EBioMedicine***54**, 102721 (2020).
21. Sandberg, Y. *et al.* TCRγδ+ large granular lymphocyte leukemias reflect the spectrum of normal antigen-selected TCRγδ+ T-cells. *Leukemia***20**, 505–513 (2006).
22. Bons, P. D. *et al.* Out of Africa by spontaneous migration waves. *PLoS One***14**, e0201998 (2019).
23. Campbell, M. C. & Tishkoff, S. A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.***9**, 403–33 (2008).
24. Pramanik, S. *et al.* Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. *BMC Genomics***12**, 78 (2011).
25. Luo, S., Yu, J. A. & Song, Y. S. Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. *PLoS Comput. Biol.***12**, e1005117 (2016).
26. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics***30**, 1312–3 (2014).

27. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.***44**, W242-5 (2016).
28. React – A JavaScript library for building user interfaces.
29. Neo4j Graph Platform – The Leader in Graph Databases.
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.***215**, 403–410 (1990).
31. Zheng, X. *et al.* Genetics and population analysis A high-performance computing toolset for relatedness and principal component analysis of SNP data. **28**, 3326–3328 (2012).

Figures

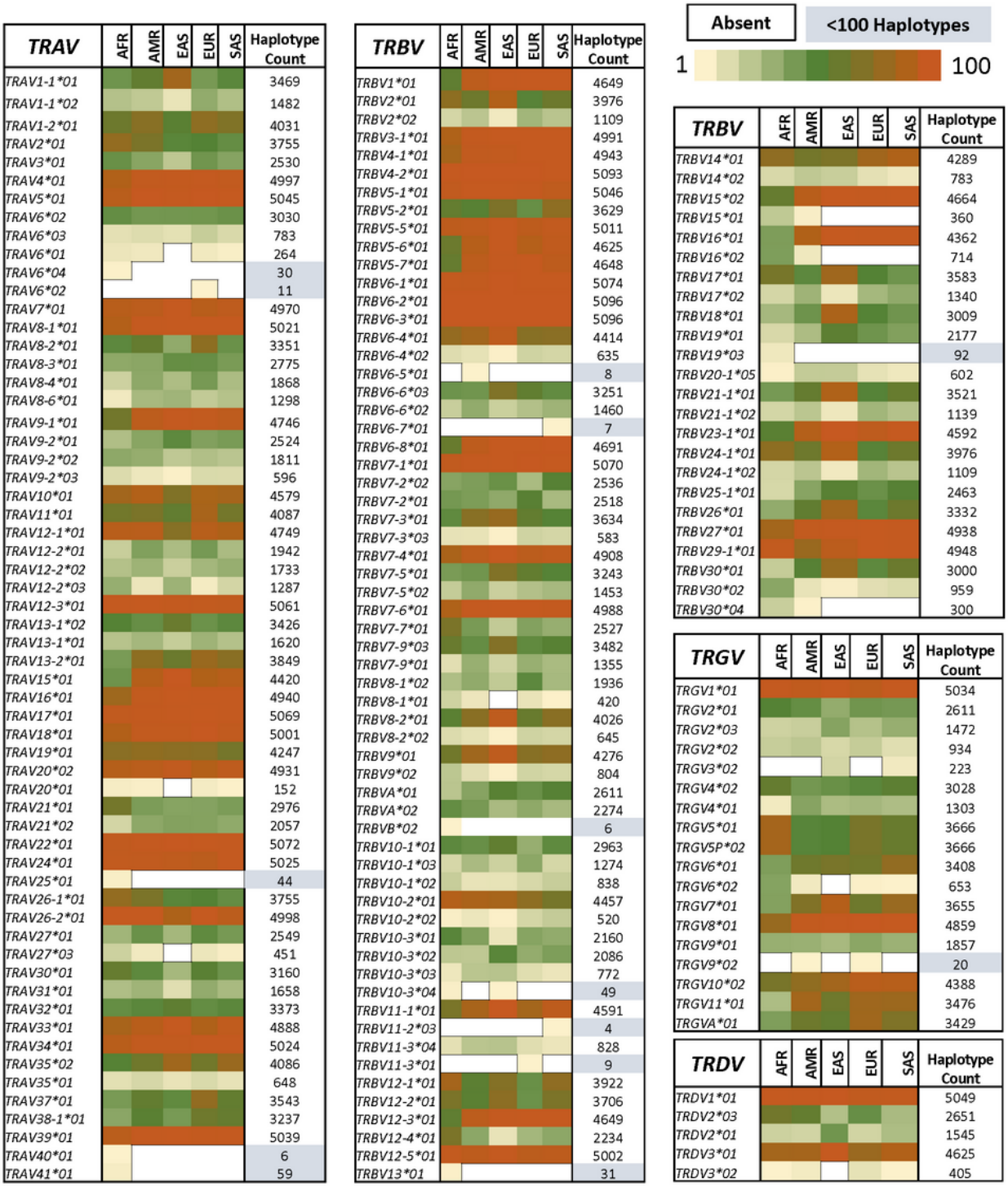

Figure 1

Heatmap representation of the super-population distribution of the known (IMGT) V gene alleles in all four human TR loci. The IMGT alleles identified in the pmTR database are shown here with the IMGT identifiers. The alleles absent in the superpopulations are not colored (white background) whereas the frequency of the alleles in the superpopulation is visualized using heatmap (yellow (0.1%) – Green (50%) – Brick Red (100%)). The last column shows the haplotype counts of these alleles. The light blue shaded

frequencies depict the alleles with <100 haplotypes. Majority of the known alleles are >30% frequent in all the super-populations.

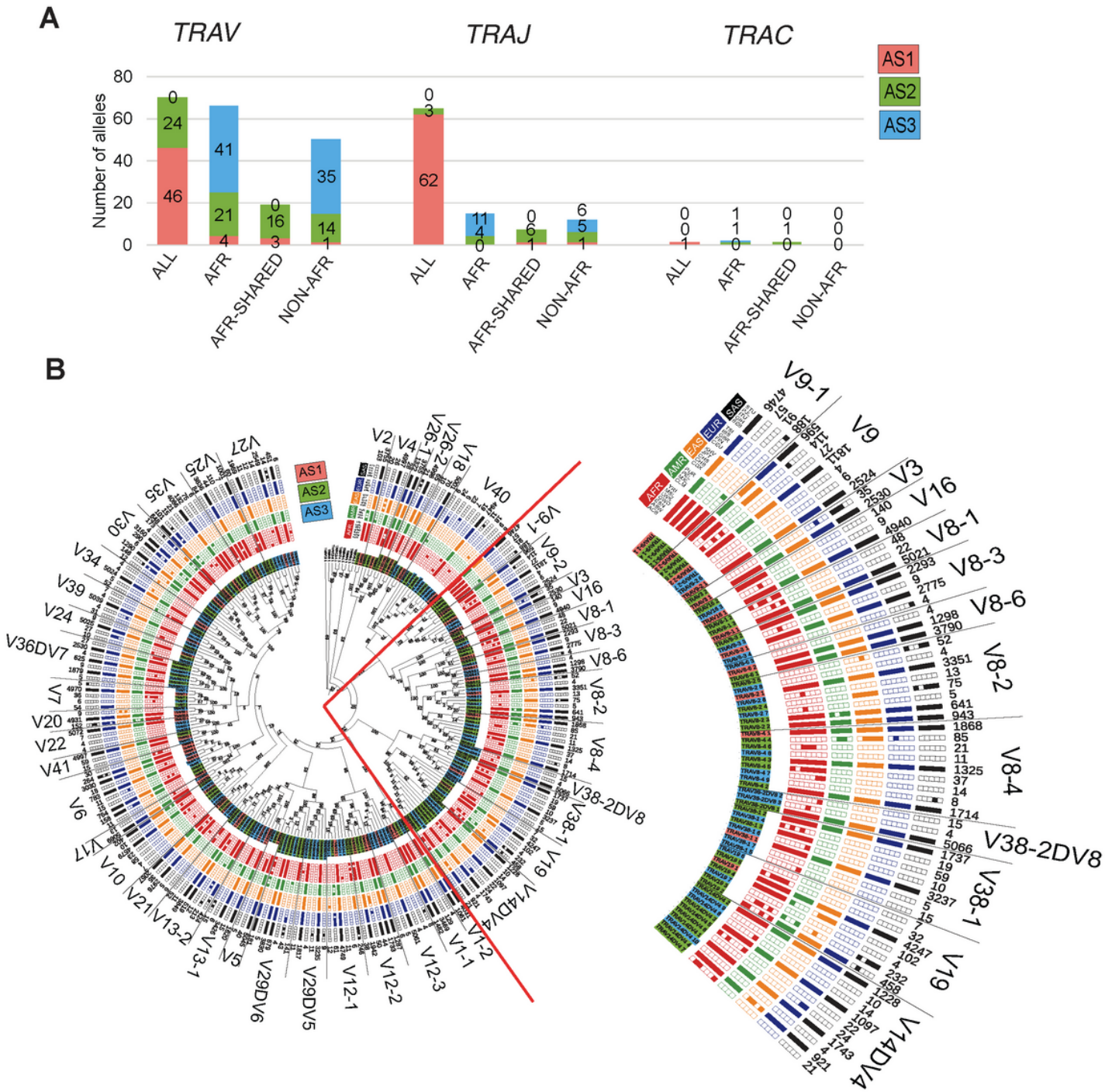

Figure 2

(Super-)population distribution of the alleles in TRA loci. A) The relative super-population distribution of VJ genes and C genes for TRA locus. The population plots are represented for all super-populations (ALL), Africans only (AFR), Africans shared with one of the other super-populations (AFR Shared) and 'Non-AFR where alleles are present in one of the populations other than Africans. Red label background indicates AS1 alleles, green AS2 and blue AS3 alleles. B) Maximum Likelihood tree of the population

distribution of TRAV alleles. Red label background indicates AS1 alleles, green AS2 and blue AS3 alleles. The population distribution is plotted in a binary format where each block is a population. Filled block represents the presence of that allele in at least four haplotypes in that population, otherwise the block is unfilled. A few TRBV alleles were used as outgroups. The bootstrap values are mentioned on the branches. The zoomed-in tree on the right to the complete tree is visualized.

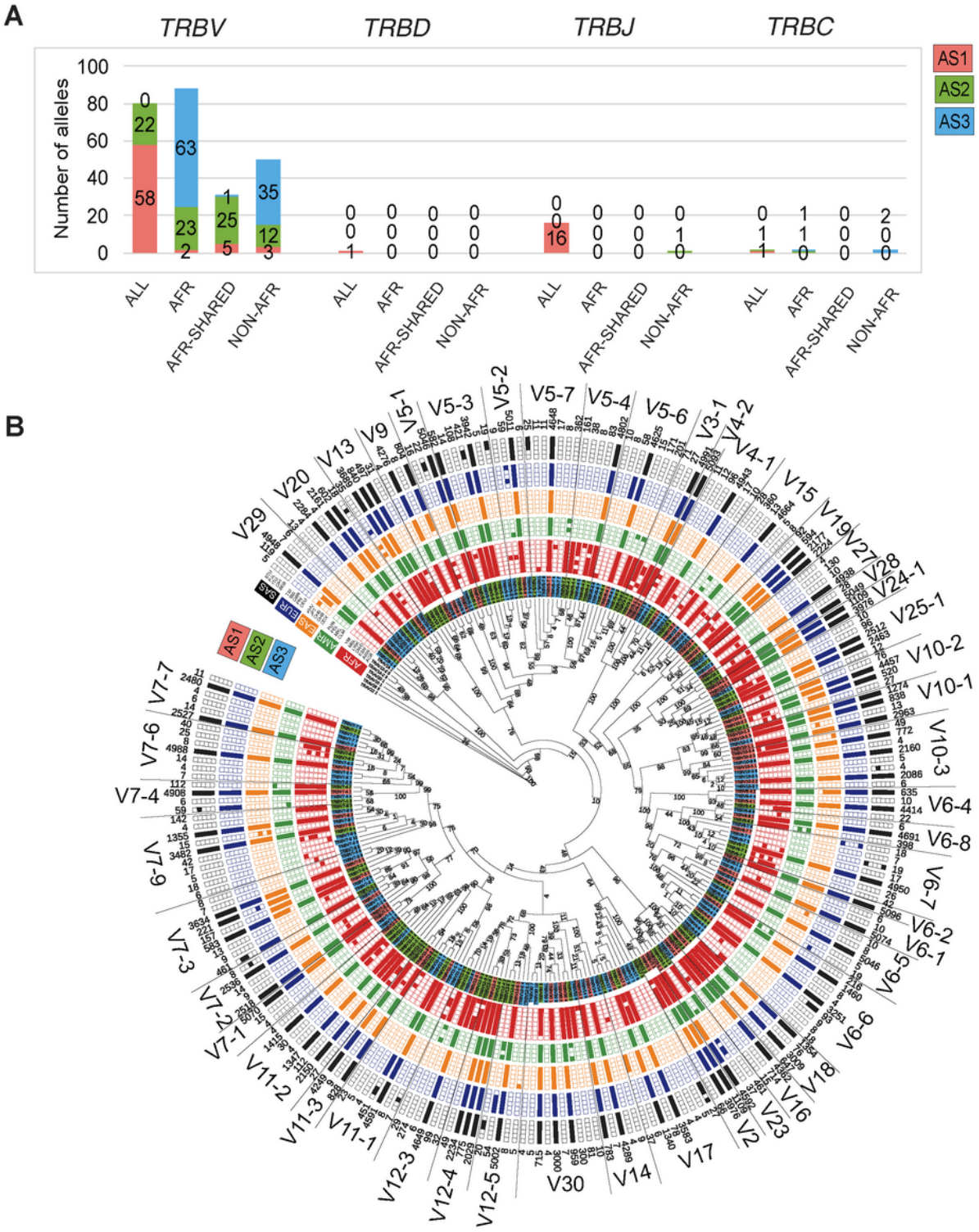

Figure 3

(Super-)population distribution of the alleles in TRB loci. A) The relative super-population distribution of VDJ genes and C genes for TRB locus. The population plots are represented for all super-populations (ALL), Africans only (AFR), Africans shared with one of the other super-populations (AFR Shared) and ‘Non-AFR where alleles are present in one of the populations other than Africans for AS1 (red), AS2 (green) and AS3 (blue) alleles. B) Maximum Likelihood tree of the population distribution of TRBV alleles. Red label background indicates AS1 alleles, green AS2 and blue AS3 alleles. The population distribution is plotted in a binary format where each block is a population. Filled block represents the presence of that allele in at least four haplotypes in that population, otherwise the block is unfilled. A few TRAV alleles were used as outgroups. The bootstrap values are mentioned on the branches.

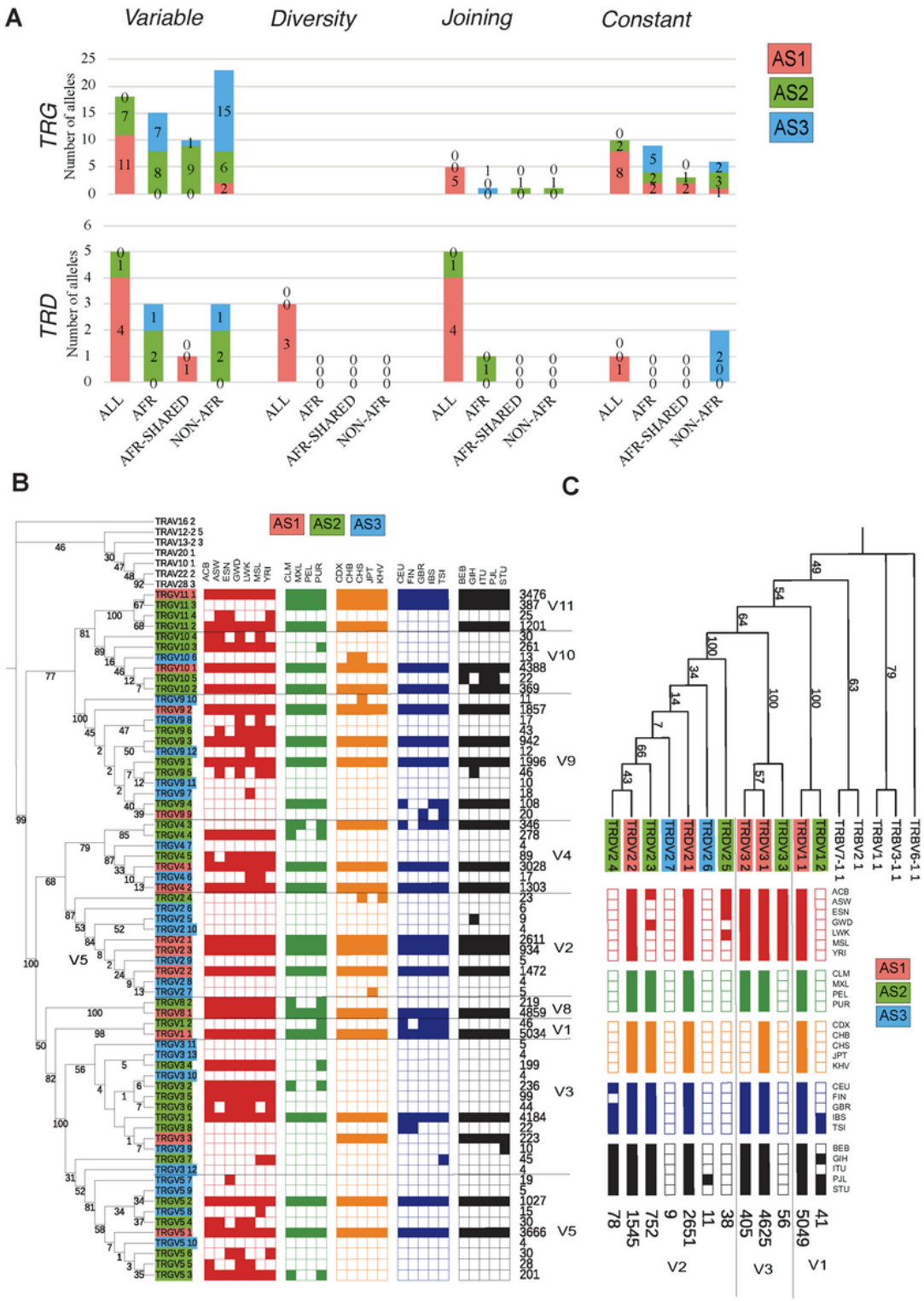

Figure 4

(Super-)population distribution of the alleles in TRG and TRD loci. A) The relative super-population distribution of VDJ genes and C genes for TRG and TRD locus. The population plots are represented for all super-populations (ALL), Africans only (AFR), Africans shared with one of the other super-populations (AFR Shared) and 'Non-AFR where alleles are present in one of the populations other than Africans for AS1 (red), AS2 (green) and AS3 (blue) alleles. B) Maximum Likelihood tree of the population distribution

of TRGV alleles. The population distribution is plotted in a binary format where each block is a population. Filled block represents the presence of that allele in at least four haplotypes in that population, otherwise the block is unfilled. A few TRAV alleles were used as outgroups. The bootstrap values are mentioned on the branches. C) Maximum Likelihood tree of the population distribution of TRDV alleles. The population distribution is plotted in a binary format where each block is a population. Filled block represents the presence of that allele in at least four haplotypes in that population, otherwise the block is unfilled. A few TRBV alleles were used as outgroups. The bootstrap values are mentioned on the branches. Please note that the nomenclature of TRDV4-8 genes correspond to TRAV genes as: TRDV4 is TRAV14, TRDV5 is TRAV29, TRDV6 is TRAV23, TRDV7 is TRAV36 and TRDV8 is TRAV38-2.

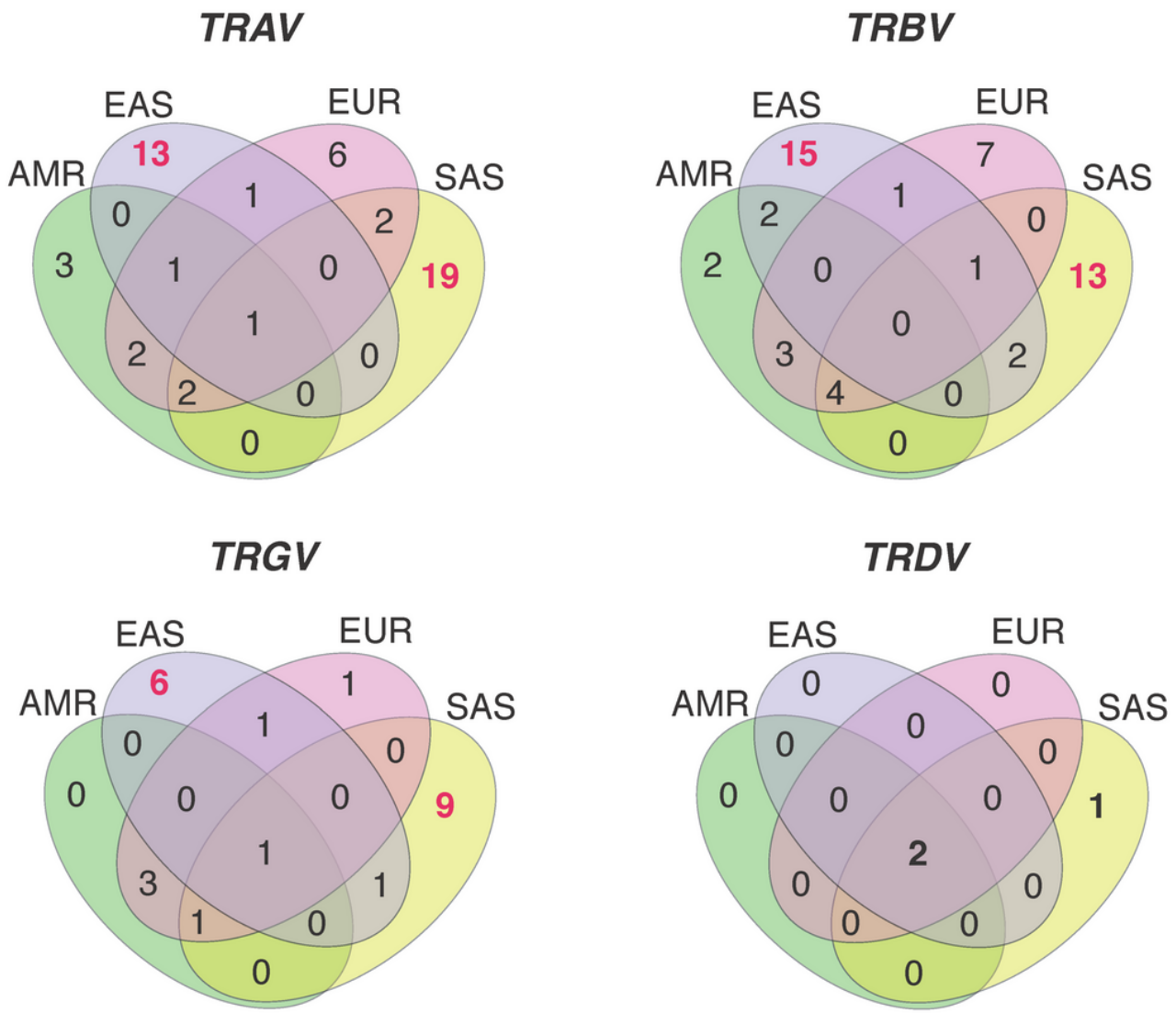

Figure 5

Super-population specificity of the Non-African alleles in TRAV, TRBV, TRGV, and TRDV genes. The number of genes specific to East and South Asian populations are shown in red color, showing that Asian

populations have more diversity

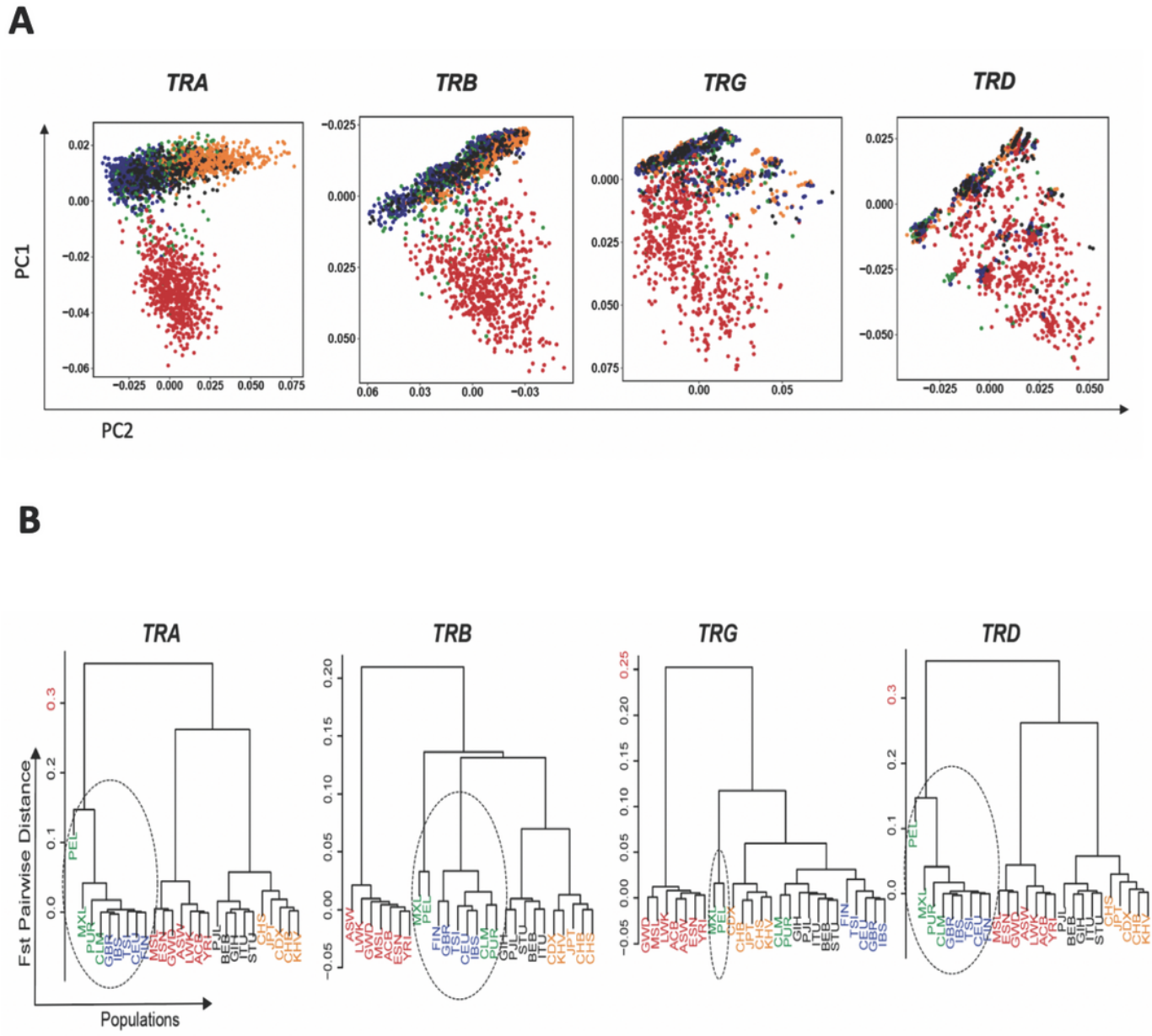

Figure 6

Genetic diversity and population structure in five super-populations for TR loci. A) Separate PCA plot of each of the four TR loci based on the polymorphisms in the complete locus. Each dot represents a sample and each sample is colored based on the super-population they belong. B) Pairwise population distribution calculated by Fst Matrix is represented as a cladogram for each locus namely TRA, TRB, TRG

and TRD. Five super-populations are colored as Africans in red; Americans in green; East Asians in orange; Europeans in blue and South Asians in black.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20210211SupplementaryFiguresmTR1.pdf
- 20210213pmTRManuscriptNSRStables.pdf