Human Immunology 80 (2019) 195–203

Contents lists available at ScienceDirect

Human Immunology

journal homepage: www.elsevier.com/locate/humimm

The variations of *TRBV* genes usages in the peripheral blood of a healthy population are associated with their evolution and single nucleotide polymorphisms

Xiao-fan Mao[a,b,1], Xiang-ping Chen[a,1], Ya-bin Jin[a], Jin-huan Cui[a], Ying-ming Pan[a], Chun-yan Lai[c], Kai-rong Lin[a], Fei Ling[b,*], Wei Luo[a,*]

[a] Clinical Research Institute, Sun Yat-Sen University Foshan Hospital, Foshan, China
[b] Department of Molecular Biology, School of Bioengineering and Biotechnology, South China University of Technology, Guangzhou, China
[c] Center of Health Management, Sun Yat-Sen University Foshan Hospital, Foshan, China

ARTICLE INFO

Keywords:
T cell receptor
Molecular evolution
Single-nucleotide polymorphisms
TRBV genes

ABSTRACT

T cell receptors (TCRs) are a class of T cell surface molecules that recognize the antigen-derived peptides presented by the major histocompatibility complex (MHC) and are able to trigger a series of immune responses. TCRs are important members of the adaptive immune system that arose in the jawed fish 500 million years ago. T cell receptor beta variable (*TRBV*) genes have been widely used to characterize TCR repertoires. Studying the evolution of *TRBV* may help us to better understand the adaptive immune system. To investigate *TRBV* evolution and its impacts on the usages of *TRBV* genes in human populations, we compared the *TRBV* genes and their homologous sequences among humans, mouse, rhesus and chimpanzee, analyzed the single-nucleotide polymorphisms (SNPs) located at *TRBV* loci, and sequenced TCR repertoires in the peripheral blood of 97 healthy donors. We found that functional *TRBV*s are more evolutionarily conserved but possess more SNPs in human populations than do nonfunctional (pseudo) *TRBV*s. Based on the conservation levels in the four species, we classified the functional *TRBV*s into 2 groups: old (conserved between mouse and humans) and new (conserved only in primates). The new *TRBV*s evolve faster and possess more SNPs than the old *TRBV*s. The variations in *TRBV* genes frequencies in the peripheral blood of healthy donors are negatively correlated with SNP density. These observations suggest that *TRBV* usages may be influenced by TCR-MHC co-evolution.

1. Introduction

T cell receptors (TCRs), a class of T cell surface molecules, primarily comprise alpha and beta chains that belong to the immunoglobulin super family. They can recognize antigen-derived peptides (8–20 amino acids) encompassed by the major histocompatibility complex (MHC) molecule and trigger a series of immune responses, including cellular activation, differentiation, proliferation, dissemination, acquisition and the deployment of effecter functions, aimed at preserving tissues and eliminating infections [2,8,9,32]. Distinct T cell clones express different TCR molecules and form a highly diverse repertoire at both the individual and population levels. TCR repertoires may be closely related to the disease status of patients. Several clinical studies had unveiled the potential utility of TCR repertoires as virus infection diagnostics and in surgery prognosis prediction and health status surveillance [2,10,39]. The TCR repertoire is highly related to human survival and worthy of in-depth investigation. Better understanding the TCR would directly promote the progress of immune therapy and benefit patients with distinct diseases. For example, the TCR gene therapy (TCR-T), a component of adoptive T cell therapy strategies, relays on rational TCR sequence modifications, which could be benefit from a better understanding of how this sequences evolve and gain important functions in human species.

A TCR is a disulfide-linked membrane-anchored heterodimeric protein normally consisting of the highly variable alpha (α) and beta (β) chains or gamma (γ) and delta (δ) chains. Each chain is composed of two extracellular domains: a Variable (V) region and a Constant (C) region. The Constant region is proximal to the cell membrane and is followed by a transmembrane region and a short cytoplasmic tail, whereas the Variable region binds to the peptide/MHC complex.

* Corresponding author.
E-mail addresses: fling@scut.edu.cn (F. Ling), luowei_421@163.com (W. Luo).
[1] These authors contributed equally.

https://doi.org/10.1016/j.humimm.2018.12.007
Received 13 May 2018; Received in revised form 10 December 2018; Accepted 17 December 2018
Available online 18 December 2018

Among all these regions, the V region of beta chains is believed to be the most specific to disease and external antigens [12]. In a human genomic context, 65 *TRBV* genes are clustered within the *TRB* locus on human chromosome 7 within 514 kbp of each other and subjected to gene recombination; during T cell development, only one *TRBV* is retained for each T cell [17,33]. The V region has three hypervariable complementarity determining regions (CDRs). CDR1 and CDR2 primarily interact with MHC helices and are entirely encoded by *TRBV* segments. CDR3 mainly interacts with antigen-derived peptides and is mainly produced by V-D-J-C segments with junction diversity.

This adaptive immune system is believed to have arisen in the jawed fishes 500 million years ago [34]. Two major events took place during the evolution of the adaptive immune system: the emergence of recombination-activating genes (*RAG*s) and massive genomic duplications [11]. The RAG-related TCR and MHC system is believed to provide an opportunity for rapid changes over evolutionary time in response to pathogens [19], whereas the genomic duplications increased the candidate gene pools for further evolution. Translocations and duplications were observed during the TCR evolution [1,40]. A common feature exists in all vertebrate *TRBV* genomic structures: the *TRBV* cluster locates at the upstream of the D-J-C repeated cluster, and a single *TRBV* locates at the end of the *TRB* locus [40]. However, *TRBV* genes varied extensively throughout vertebrate evolution and became somewhat conserved only after the divergence of primate lineages (65 million years ago) [25].

According to the functional annotations from IMGT (www.imgt.org), human *TRBV*s are classified into 3 types: pseudo, open reading frame (ORF) and functional. Twelve human *TRBV*s have been confirmed as pseudo genes; seven have been identified as ORFs, and forty-seven have been identified as functional genes. A recent study explored the differences in amino acid composition between pseudo and functional *TRBV*s [37], and the results suggest that distinct conservative patterns exist in these two groups: Some amino acids were more frequently used in pseudo *TRBV*s, whereas others were not. The conserved patterns of amino acid composition are mostly the consequence of evolution. However, it remains unclear how evolution has influenced population polymorphisms and whether it has influenced the usages of *TRBV*s in individual humans. Addressing these questions can help us understand the formation and evolution of the human immune system.

To this end, we analyzed human *TRBV* evolutionary features (evolutionary rates, sequence divergence and phylogenetic trees) using three comparison species, used the SNPs in *TRBV*s to identify *TRBV* polymorphisms in a human population, and identified human *TRBV* usages in peripheral blood by sampling 97 healthy donors. Association analyses were performed to unveil the correlations among the evolution features, polymorphisms and usages of human *TRBV*s. This study provides new insights into the evolution of the adaptive immune system.

2. Materials and methods

2.1. DNA sequences from public databases

The reference genome sequences were all downloaded from UCSC Genome Browser (https://genome.ucsc.edu/). The genome versions were as follows: GRCh38 for humans, rheMac8 for rhesus, panTro5 for chimpanzees, and GRCm38 for mouse. The human *TRB* locus nucleotide sequence was obtained from NCBI GenBank (GeneID: 6957), which can be completely mapped onto chromosome 7 of the human reference genome.

2.2. Genomic coordinates and nucleotide sequences of human gene sequences and their homologous sequences in the comparison species

Human *TRBV* gene coordinates were obtained from the annotation by NCBI GenBank. Based on these coordinates and using the UCSC LiftOver tool from UCSC (http://genome.ucsc.edu/cgi-bin/hgLiftOver), we located the homologous sequences in the comparison species, and we retrieved the nucleotide sequences according to the sequence locations on the reference genome. In total, 61 TRBVs were located in the human genome.

2.3. Determination of functional and nonfunctional (pseudo) TRBVs

Human *TRBV* nucleotide sequences and their homologous sequences in the comparison species were translated into amino sequences utilizing the function “Codon” of BioPerl (http://bioperl.org/). Human *TRBV* genes and their homologous sequences in the comparison species were defined as pseudo or nonfunctional genes only when inappropriate positions of start/stop codons were observed. The ORF *TRBV*s were grouped into the functional group since the functionalities of ORFs were unclear. The annotations were verified in the IMGT database (http://www.imgt.org/) to ensure all pseudo TRBVs were correctly annotated.

2.4. TRBV sequence alignments between human and comparison species, phylogenetic tree construction, and Ka/Ks calculation

TRBV coding sequences and protein sequences from human and comparison species were aligned using ParaAT [44] and T-Coffee [27]. Trees for each *TRBV* were constructed using MEGA X [20] based on the nucleotide alignment results. The Ka/Ks (nonsynonymous substitution rate / synonymous substitution rate) values for *TRBV*s were calculated using the ‘codeml’ method of PAML software [43]. Three models were performed including the branch models (model = 1, NSsties = 0), site models (model = 0, NSsties = 0 1 2 3 4 5 6 7 8) and branch-site models (model = 2, NSsties = 2, branches of non-primates and primates were labeled respectively). Likelihood ratio tests (LRT) were performed between the M2 and M1 as well as M8 and M7 results of site models respectively to test the significance of positive selections [4]. Bayes Empirical Bayes (BEB) analysis was used to detect the positive selection sites. All scripts, parameters and results for the alignments, tree constructions and Ka/Ks calculations are provided in Supplementary file 1. Full length human *TRB* and the homologous sequence in mouse, rhesus and chimpanzee were aligned and the dot plots generated by DOTMATCHER from EMBOSS [31] were used to exhibit the similarities between each two sequences. The window size and identity threshold were set to 500 and 75 when comparing the rhesus and chimpanzee to humans. When comparing mouse to humans, we set the window size to 200 and the identity threshold to 50 since TRB sequences of these two species were highly diverse. Based on the TRBV protein sequence alignment results from T-Coffee, a phylogenic tree was generated by MEGA X using the neighbor-joining method.

2.5. Calculation of SNP densities of TRBV genes

SNP data were downloaded from dbSNP150 [36]. All SNPs located in the *TRB locus* were retrieved. For each *TRBV*, the SNP density was calculated as the total number of SNPs located in the *TRBV* coding region divided by the *TRBV* coding sequence length.

2.6. Sample collection

All of the clinical samples in this study were provided by Sun Yat-Sen University Foshan Hospital (Guangdong, China). Peripheral blood mononuclear cells (PBMCs) from 97 healthy donors were collected. All donors were diagnosed “Normal” through medical examinations. Information on all of the donors is provided in Supplementary file 2. All fresh PBMCs were lysed with TRIzol® Reagent (Invitrogen, USA) and frozen at −80 °C until further processing. This study was approved by Sun Yat-Sen University Foshan Hospital. All donors provided written informed consent, and all experiments were conducted according to the guidelines of the Declaration of Helsinki.

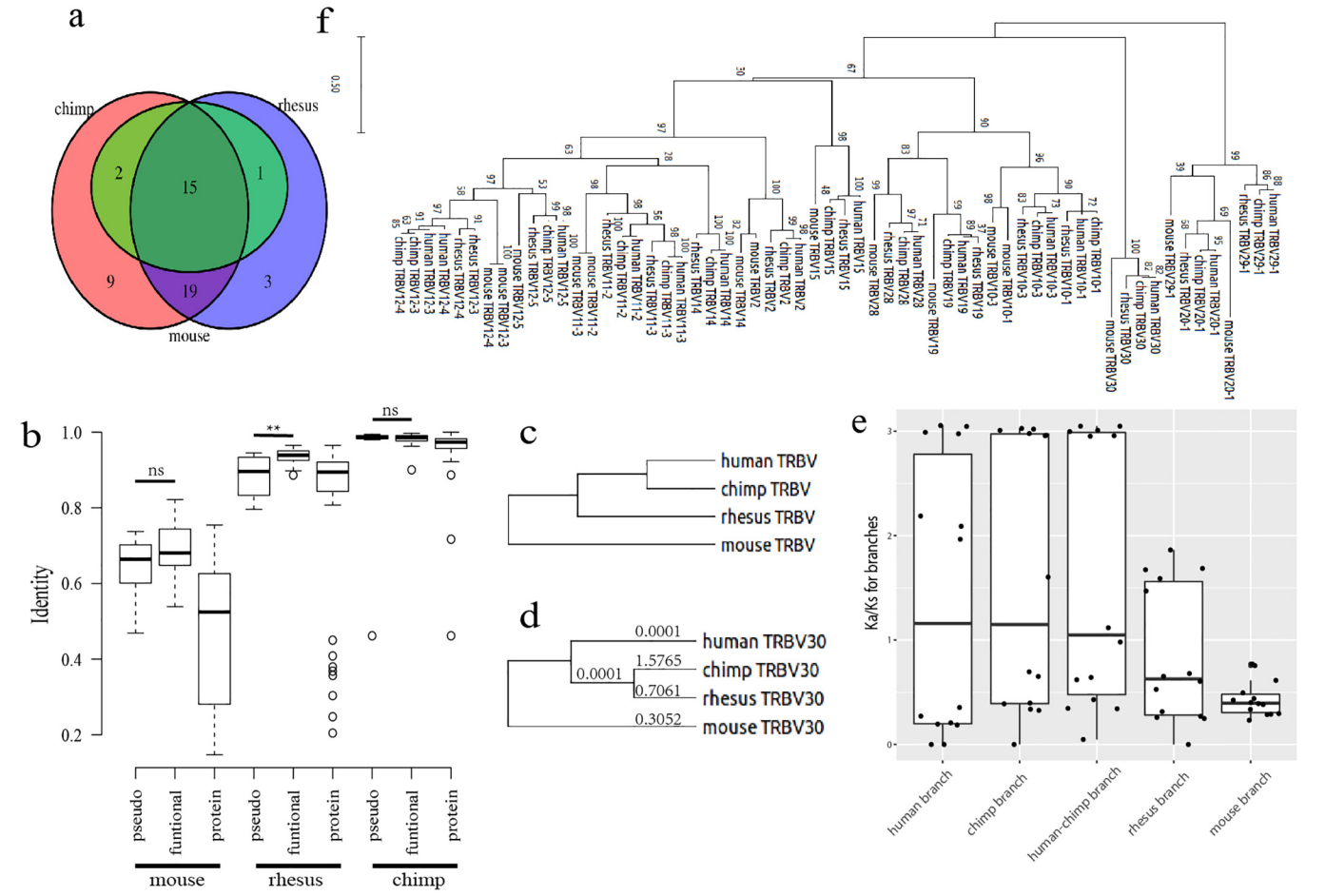

Fig. 1. Homologous *TRBV* genes between humans and three model species. a) Venn plot of the overlap of human functional *TRBVs* with homologous sequences that were also functional in the three comparison species. b) Sequence identities were calculated by comparing the human *TRBV* genes and protein sequences to their homologous sequences in mouse, rhesus and chimpanzee. *TRBV* genes were divided into a pseudo group (pseudo) and a functional group (functional). Protein sequences (protein) used for the alignments were translated from the functional human *TRBV* genes and their homologous sequences. c) and d) Phylogenetic trees were constructed for all conserved 15 *TRBVs* from Fig. 1a (the intersection of rhesus, chimpanzee and mouse) based on the alignments of nucleotide sequences in the four species. The tree topologies were the same for all *TRBVs* (Fig. 1c) except for *TRBV30* (Fig. 1d). These trees (Fig. 1c&d) were used in later Ka/Ks analysis using the 'codeml' function in PAML. Along with each branch of the tree (Fig. 1d) are the Ka/Ks values for *TRBV30*. e) Ka/Ks values for the other *TRBVs* (the conserved 15 *TRBVs* minus *TRBV30*, n = 14) in all branches are shown using the branch models in PAML. f) An evolutionary tree constructed based on TRBV protein sequences (n = 15) that are functional in human, mouse, rhesus and chimpanzee. The percentage of trees in which associated taxa clustered together is shown next to the branches. Branch length represents the number of substitutions per site.

2.7. TCR sequencing preparation

Total RNA was extracted from 1 ml of PBMC lysate using a total RNA Kit (OMEGA, USA) according to the manufacturer's instructions. Approximately 1 μg of total RNA was reverse transcribed into 5′ RACE-ready cDNA by using the SMARTer PCR cDNA synthesis kit (Clontech, USA). To obtain specific, clean products, a TCR library for sequencing was prepared through semi-nested PCR amplifications. The first-round reaction was performed using 1 μg cDNA and 0.2 ml of Advantage 2 polymerase mix (Clontech, USA), Nested Universal Primer (NUP, Clontech, 5′-AAGCAGTGGTATCAACGCAGAGT-3′) and 3′-TCRβ outer primer (5′-AGATCTCTGCTTCTGATGGCT-3′). The first-round PCR products were purified with a gel extraction kit (QIAGEN, German) and used as template for the second amplification with 3′-TCRβ inner primer (5′-TGGCTCAAACACAGCGACCT-3′). The 3′-TCR β outer and inner primers were homologous to both *TRBC1* and *TRBC2* and were validated as reliable nested primers for TCR amplification in our previous study [24].

2.8. RNA sequencing and analysis

The *TRB* libraries were sequenced (2 × 150 bp) using the Illumina HiSeq 2500 platform. Low-quality sequences were discarded, and sequence reads beginning with *TRBC1* and *TRBC2* sequences were extracted. BLAT [18] was used to identify *TRBV* genes of every read based on the reference *TRB* sequences (GenBank GeneID: 6957). Nucleotide sequences were translated into amino acid (aa) sequences, which were aligned to find specific *TRBVs*. A subset of these alignment sequences without V -J-C genes was then filtered. A one-step tool was developed for generating the TCR repertoire data from the Illumina sequencing data. The tool can be downloaded from the website along with one example: https://github.com/jinyabin1990/TCR_one_step.

2.9. Population distribution indexes of TRBVs and statistical analysis

Each *TRBV* usage in the population was estimated by its frequency, which was calculated as the number of mapped read counts that mapped to the *TRBV* divided by the total read counts that mapped to the total *TRBVs*. We used three equality/inequality measures to estimate the variation in *TRBV* usage in the peripheral blood of healthy

donors.

Shannon entropy: $SA = \sum_{i=1}^{n} (x_i) \ln(x_i)$

Relative standard deviation: $RSD = \sqrt{\sum_{i=1}^{N} (x_M - x_i)^2} / (n-1)x_M$

Gini coefficient: $Gini = \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| / \left(2n \sum_{i=1}^{n} x_i \right)$

For the formulas above, x_i and x_j (where i and j range from 1 to 97) refer to the *TRBV* frequency of the related donor, x_M is the mean *TRBV* frequency of all donors, and n is the total number of donors. We did not calculate the diversity index of the TCR repertoire for each donor; rather, we calculated the indexes for each *TRBV* to estimate the usage deviations among different donors, to see if one TRBV is stable among different individuals in population.

A heatmap was produced and hierarchical clustering of *TRBV*s was performed in R using the 'pheatmap' package. All statistical analyses were performed in R.

3. Results

3.1. Comparisons of human TRBV genes with the TRBV genes of three other species

To investigate how TCR sequences evolved in mammalian species, we used the mouse, rhesus and chimpanzee as the comparison species for human. Over 90% of the human *TRBV*s were mapped to homologous sequences in these species, using the LiftOver tool from the UCSC. Both coding sequences and the protein sequences were used to perform comparisons.

Human *TRBV* genes were divided into two groups: 1) functional *TRBV*s, which potentially express functional protein sequences (n = 49) and 2) pseudo *TRBV*s, which do not produce functional proteins (n = 12). A *TRBV* was defined as a pseudo *TRBV* in our study only if inappropriate start/end codons were observed, which was consistent with the annotations from the IMGT (Section 2.3). The overlap of functional human *TRBV*s with functional homologous sequences in the comparison species is shown in a Venn plot (Fig. 1a). All human functional *TRBV*s were mapped to homologous sequences in at least one comparison species. A phylogenetic tree (Fig. 1f, Section 2.4) was generated based on the functional and homologous TRBV protein sequences (n = 15 * 4) in all four species to exhibit the sequence divergences. For the most part homologous genes in the different species cluster together in the tree. However, there are a number of exceptions, where different TRBV from the mouse cluster closer together, such as mouse TRBV10-1 and mouse TRBV10-3. This suggested that mouse TRB locus was still undergoing duplications which are the major way that *TRBV* gene expanded.

DNA and protein sequence identities were calculated from the alignments between the human *TRBV*s and the homologous sequences in the comparison species (Fig. 1b). Not unexpectedly, the human *TRBV* sequences were most conserved in chimpanzee and least in mouse, which is consistent with the phylogenetic tree results. The protein sequence identities were lower than the nucleotide sequence identities, suggesting that nonsynonymous substitutions and reading frame shifts exist between the human and comparison species.

For each conserved *TRBV* (n = 15), we constructed an unrooted phylogenic tree based on the nucleotide sequences in the four species. Except for *TRBV30* (Fig. 1d), all trees exhibited similar topology (Fig. 1c). The evolutionary rates (Ka/Ks values) were calculated for these *TRBV*s using the 'codeml' function in PAML based on the related tree topology. First, we performed the branch model which allows the Ka/Ks to vary among branches in the phylogeny and are useful for detecting positive selection acting on particular lineages [26]. The Ka/Ks value of each branch is shown in Fig. 1e (*TRBV30* was excluded for its phylogenetic tree was different form the others). This result suggested that some of the conserved *TRBV* genes had undergone accelerated evolution, especially in the human-chimpanzee branch. However the significances of the positive selections were not given in this model. Secondly, we performed the site models which allow the Ka/Ks to vary among codons. The results showed that the average Ka/Ks were below 1 for most *TRBV*s, which is not surprising since positive selection is unlikely to affect all sites. Our results also showed that only 3 *TRBV*s possessed one positive selection coden, while others did not possess any positive selection site. The LRT tests between M7 (beta) and M8 (beta&ω) showed that only *TRBV20-1* and *TRBV28* might under positive selection (LRT, $p < 0.05$). A result summary of site models was provided in Supplementary file 4. Finally we conducted the branch-site model, which allow Ka/Ks to vary both among sites and across branches on the tree to see if positive selection sites may exist between the primate lineage and non-primate lineage. We found that 3 sites of *TRBV2* and 1 site of *TRBV20-1* were under positive selection (BEB, $p < 0.05$). All analysis showed that a small section of these *TRBV*s were under positive selection especially in the primate lineage, while others were not, which is comprehensive since these 15 *TRBV*s are the most conserved *TRBV*s across human, rhesus, chimpanzee and mouse. All PAML results were provided in Supplementary file 1.

3.2. Human TRB locus VS the homologous gene in comparative species

To investigate how the entire TRB loci evolved among these three species, we performed global alignments among these three species.

We aligned human *TRBV* genes to the homologous sequences in the genome of all three species by using the UCSC LiftOver tool (Sections 2.2 and 2.4). We found that the positions of homologous *TRBV*s in rhesus and chimpanzee were mostly identical to the corresponding positions in humans (Fig. 2b&c). One duplication and one translocation were observed between human and rhesus (Fig. 2b). No duplication and translocation was observed between humans and chimpanzee (Fig. 2c). However, homologous *TRBV*s were not so identical between humans and mouse, and we could observe many duplications between human and mouse *TRB* sequences (Fig. 2a). These results revealed a conserved pattern of *TRBV* gene position among the primates, suggesting high conservation of the recognition mechanism in V-D-J somatic recombination in primates.

Based on the conserved sequence structure between the human *TRB locus* and its homologous sequence in rhesus, we performed a full-length alignment to globally review the *TRB* locus in humans and other species (Section 2.4). The alignment is shown in a dot plot (Fig. 2d–f). *TRB* homologous sequence in mouse showed significant difference to the human TRB sequence. Two major duplication regions were observed in rhesus and chimpanzee. Some duplicated sites were deleted in humans. Thus, *TRBV*s were closer to each other in humans than in rhesus and chimpanzee, and each human *TRBV* might be more evenly and efficiently selected during T cell maturation, which might contribute to the high diversity of T cell repertoires in humans. This possibility could be examined by analyzing the correlation between *TRB* sequence length and TCR diversity in multiple species in future studies.

All of these results demonstrated that although most of the human *TRBV* genes could be mapped in three comparison mammal species, there were significant differences between the primates and non-primates. These results are consistent with a previous conclusion that the TCR repertoire was very stable only in primates [25,38]. Additionally, some *TRBV* genes are exhibiting accelerated evolutionary rates within the primates, especially in the human-chimpanzee branch.

3.3. The SNPs in human TRBVs

We downloaded dbSNP 150 and associated the SNPs to the human

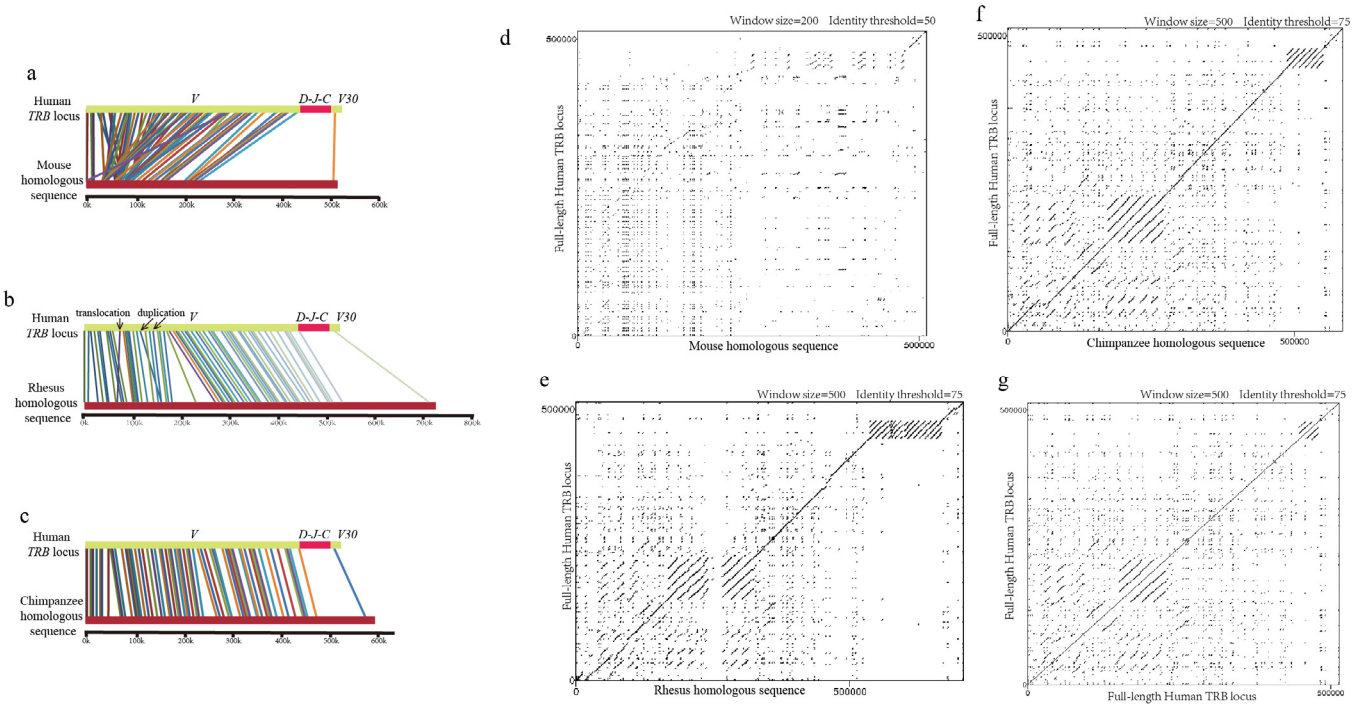

Fig. 2. The alignment results of homologous *TRB* locus between humans versus mouse, rhesus and chimpanzee. The loci of homologous *TRBV*s were matched between human versus mouse (a), rhesus (b) and chimpanzee (c). The lines connected homologous *TRBV*s between different species. Human *TRBV* regions were marked in green. D-J-C referred the *TRBD-J-C* regions in human *TRB* and was marked in red. V30 referred to the *TRBV30* gene at the end of the *TRB* locus. b) The dot plot of full-length alignments of homologous *TRB* locus between humans versus mouse (d), rhesus (e) and chimpanzee (f) were generated by DOTMATCHER in EMBOSS (Section 2.4). g) Self-alignment of the human TRB locus was also performed to detect the duplication sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

TRBV locations to investigate whether SNP density was related to the evolutionary characteristics of *TRBV* genes.

The human functional *TRBV*s were divided into 2 groups: old and new. The old group represented the human *TRBV*s that are also functional in mouse (n = 1 + 15 + 2, Fig. 1a), whereas the new group represented the human homologous *TRBV*s that are not functional in mouse but are functional in rhesus or chimpanzee (n = 9 + 19 + 3, Fig. 1a). Thus, all *TRBV*s were classified into 3 groups: 1) pseudo genes (n = 12), 2) old genes (n = 18) and 3) new genes (n = 31). The grouping details are listed in Supplementary file 3. Comparison of homologous *TRBV*s between rhesus and human revealed that the old *TRBV*s were the most conserved type, whereas the pseudo *TRBV*s were the least conserved (Fig. 3a). Furthermore, the new *TRBV*s possessed higher evolutionary rates (Ka/Ks values) than did the old *TRBV*s when only used rhesus as comparative species (Fig. 3b). This finding is consistent with the previously proposed hypothesis that a new gene in its early stage typically undergoes rapid changes in sequence, structure and expression, which indicates a continuous evolution of function; a significant role of positive selection in these changes can often be detected [23].

Another useful indicator of natural selection is the abundance of polymorphisms [14]. Thus, we downloaded the SNP data from dbSNP and retrieved all SNPs that are located at the TRB locus to observe *TRBV* polymorphisms. We identified 57 k SNPs located at the locus of TRB (514 k). The average SNP density was 0.1109 (57 k/514 k) for the TRB locus. When associating the SNPs to the related *TRBV*s, we found that SNP densities were the lowest in pseudo *TRBV*s and the highest in new *TRBV*s (Fig. 3c, Wilcoxon rank sum test, pseudo VS new: $p = 0.002317$; new VS old: $p = 0.01438$; pseudo VS functional: $p = 0.01879$). More conserved genes can be expected to possess lower levels of sequence polymorphism than do less conserved genes [14]. Accordingly, in this study, the more conserved, old *TRBV*s possessed fewer SNPs than did the less conserved, new *TRBV*s. However, compared to the pseudo *TRBV*s, the old and new *TRBV*s are more conserved (Fig. 3a) but have higher numbers of SNPs (Fig. 3c). This observation may be attributed to the TCR-MHC restriction that the usage of functional *TRBV*s are restricted by MHC molecular type [3]. The pseudo *TRBV*s are not restricted by the MHC as they are their nonfunctional and are under neutral evolution. Thus, the SNP densities of pseudo *TRBV*s were close to the average SNP density at the *TRB* locus (Fig. 3c, the *TRB* average SNP density is marked by a red line), in which most sites were noncoding regions and under neutral selections. The new *TRBV*s had the highest number of SNPs (Fig. 3c), suggesting that the new *TRBV*s may have tightly co-evolved with the highly diverse MHC molecules.

We observed a significant difference in SNP density of the *TRBV*s between the 5′-half of *TRB* (n = 30) and the 3′-half (n = 31) (Fig. 3d&f, Wilcoxon rank sum test, $p = 0.000775$). We calculated the DNA sequence substitution rates between humans and rhesus for the *TRBV*s at the 5′- and 3′-halves. The results showed that the DNA sequence substitution rates in *TRBV*s at the 5′-half were significantly higher than those in *TRBV*s at the 3′-half (Fig. 3e, Wilcoxon rank sum test, $p = 0.005547$), which is consistent with the evolutionary signature of more conserved genes (genes with fewer substitutions) possessing fewer SNPs. Additionally, 13 of 18 (72.2%) old *TRBV*s were located in the 3′-half (Fig. 3f). These results demonstrated that evolutionary forces were unevenly distributed throughout the length of the *TRB* locus. The 3′-half was more conserved than the 5′-half. We consider this may be related to the evolutionary patterns of the whole TRB locus, where duplications were mostly observed in the 5′-half of TRB locus from the results of full-length TRB locus alignment (Fig. 2e–g).

3.4. The pseudo, old and new TRBV usage distributions in the peripheral blood of 97 healthy donors

*TRBV*s in different groups show distinct evolutionary patterns, and *TRBV*s exhibit different levels of polymorphisms among populations.

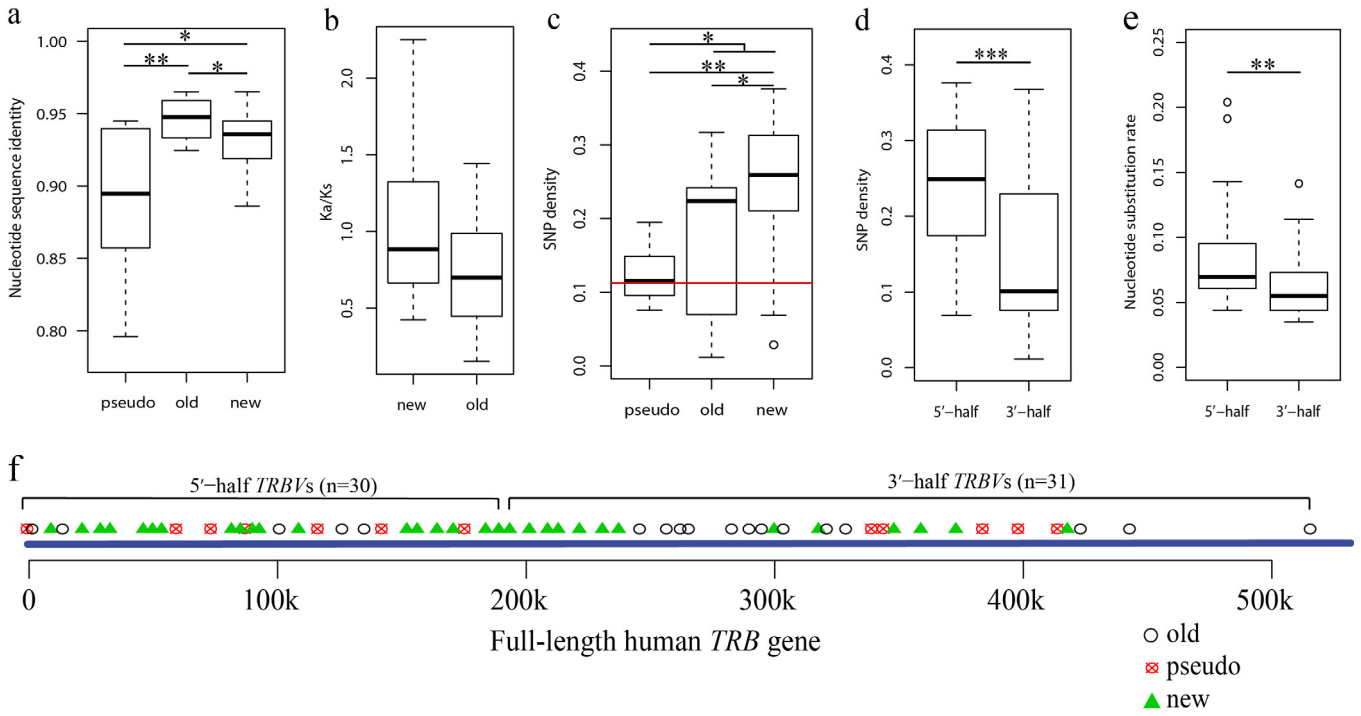

Fig. 3. Comparisons of pseudo, old and new *TRBV*s. a) Nucleotide sequence identities of the pseudo, old and new *TRBV*s between human and rhesus.. b) The evolutionary rates (Ka/Ks values) of the new and old *TRBV*s using rhesus as the comparison species. c) The SNP densities of the pseudo, old and new *TRBV*s. The average SNP density in the full-length TRB locus is represented by the red line. d) The SNP densities of the 5′-half and 3′-half *TRBV*s. e) The substitution rates of the 5′-half and 3′-half *TRBV*s. The substitution rates are calculated based on the gene sequence variations between humans and rhesus. f) The *TRBV* gene loci on the human *TRB* locus. The pseudo, old and new *TRBV*s are marked along the full length of the human *TRB* locus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We were interested in determining how the usage of the three groups of *TRBV*s is distributed in the peripheral blood of human populations. Thus, we performed RNA sequencing of the peripheral blood TCR repertoires of 97 healthy donors aged 19 to 70 (Supplementary file 2).

It has been reported that TCR repertoire diversity is negatively correlated with age [13]. Consistently in our cohort, the values of Shannon entropy, a diversity index, of all donor TCR repertoires in PBMCs were negatively correlated with age (Pearson correlation test, $p = 0.000572$). However if *TRBV* usages were influenced by age, our results would be influenced by the sampling the cohort with different ranging of age. To evaluate whether some *TRBV* usages may be influenced by age, we performed multiple Pearson's correlation tests between usage of each *TRBV* and donor age, with p values adjusted by the Benjamini and Hochberg method. No correlation was observed between any *TRBV* and age, suggesting that the age did not significantly influence *TRBV* usage.

The *TRBV* usages of each donor are shown in Fig. 4a. In general, the gene usages varied a lot among different *TRBV*s. However, the variations of *TRBV* usages between individuals were low. Besides, we found that pseudo *TRBV* usages were significantly lower than were old *TRBV* usages (Wilcoxon rank sum test, $p = 2.11e-5$, Fig. 4b) and that they could be hierarchically clustered (Fig. 4a). However, usages of the new *TRBV*s did not significantly differ from those of old *TRBV*s (Fig. 4b). We also investigated the population distributions of the three groups based on *TRBV* Gini coefficients, Shannon entropies and relative standard deviations (Section 2.9, Fig. 4c–e). The results showed that the usages of pseudo *TRBV*s were more diverse than those of the other *TRBV*s in human populations, with usages of the old and new *TRBV*s being less variable.

3.5. Variation in TRBV gene usage was negatively correlated with SNP density

To understand how *TRBV* polymorphisms are associated with *TRBV* usage, we performed Pearson correlation tests of the *TRBV* distribution indexes (Section 2.9) and the SNP densities at *TRBV* loci. In our cohort, we did not observed any correlation between the frequency of *TRBV* usage and SNP density (Fig. 5a, Pearson correlation test $p = 0.81$). However, we observed significant correlations between *TRBV* SNP density and each of the *TRBV* usage variation indexes (Fig. 5b–d), suggesting that the usage of *TRBV*s with higher SNP densities is more stable among populations than is that of *TRBV*s with lower SNP densities. When considering only new *TRBV*s, similar correlations were observed, with SNP density being negatively correlated with Gini coefficient (Pearson correlation test, $p = 0.0438$, not shown in figure), positively correlated with Shannon entropy (Pearson correlation test, $p = 0.01027$, not shown in figure) and negatively correlated with relative standard deviation (Pearson correlation test, $p = 0.01137$, not shown in figure). Hence, we confirmed that the *TRBV* distributions in the peripheral blood of a healthy population were correlated with the SNP densities of the related *TRBV* coding regions.

This relationship might also yield the interesting topic of TCR-MHC co-evolution. Some recent studies had confirmed that there exist TCR germ line biases for the MHC molecule indicating a co-evolving pattern of TCR and MHC. For example, Sharon *et al.* showed linkage between multiple MHC and TCR genes using expression quantitative trait loci analysis [35]. Blevins et al. supported an evolved compatibility exists alongside MHC bias in the human TCR repertoire [5]. We considered our results in agreement with these findings since the SNPs in *TRBV*s may influence their usages. *TRBV*s with higher SNP may co-evolve with MHC more closely, thus lead to a more stable usages in population.

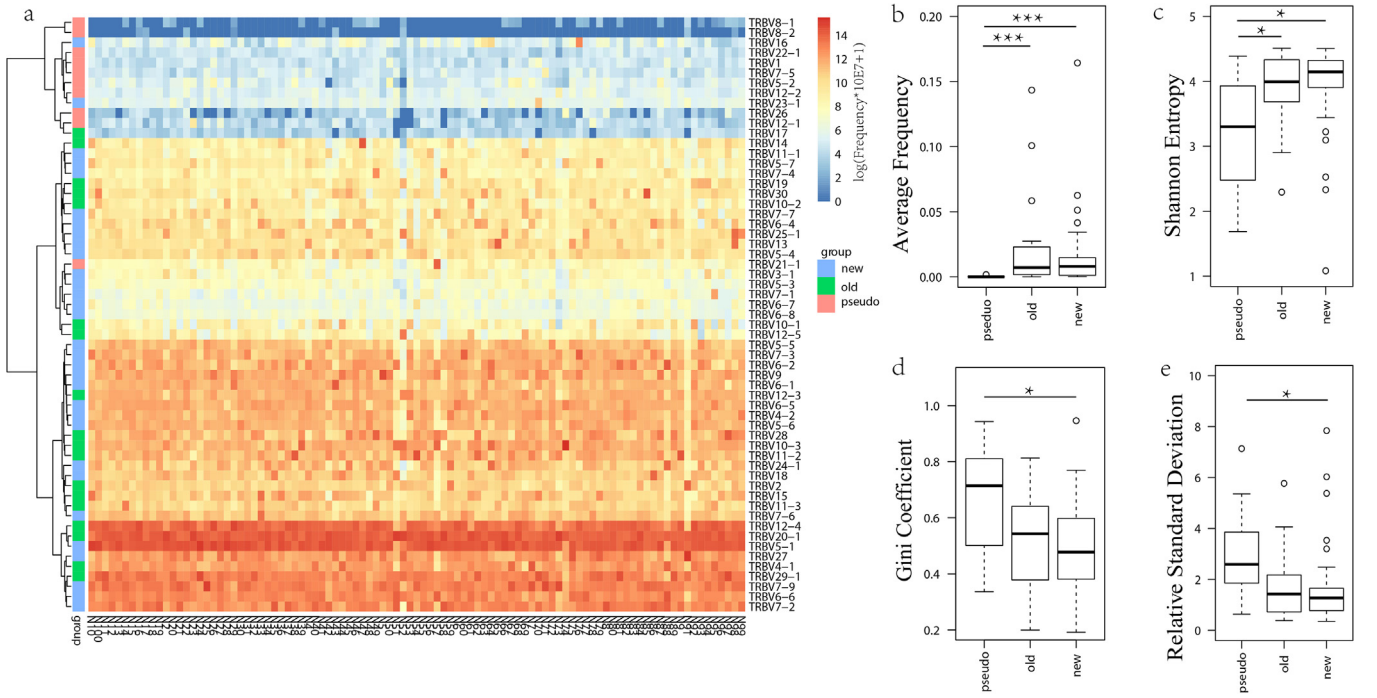

Fig. 4. *TRBV*s usages and their distribution indexes. For each *TRBV*, the average frequency of its usage in all samples, Shannon entropy, Gini coefficient and relative standard deviation were calculated based on the *TRBV* usages in all donors (Section 2.9). a) *TRBV* usages (in rows) in peripheral blood of 97 healthy donors (in columns). *TRBV*s are hierarchically clustered on the left side with group annotations. b) Box plot of the average frequencies of pseudo, old and new *TRBV*s. c) Box plot of Shannon entropy values of the pseudo, old and new *TRBV*s. d) Box plot of Gini coefficients of *TRBV*s in each group. e) Box plot of relative standard deviations of *TRBV*s in each group. * referred to the Wilcoxon rank sum test p-value lower than 0.05; ** referred to the Wilcoxon rank sum test p-value lower than 0.01; *** referred to the Wilcoxon rank sum test p-value lower than 0.001.

4. Discussion

Natural selection is one of the strongest forces driving species evolution. Positive and negative selections help purify detrimental mutations and promote beneficial mutations during a species' evolution. Typically, conserved genes are old and slowly evolving, possess low SNP densities, have high and stable expression, are subjected to negative selection and perform important functions [16,21,22,41,42]. In contrast, non-conserved genes exhibit the opposite properties.

However, our study showed that the evolution of *TRBV* genes does not conform to these patterns. The SNP densities in pseudo *TRBV*s are similar to the average SNP density of full-length TRBs, in which most sites are noncoding regions, suggesting that pseudo *TRBV*s are under neutral selection. This is not surprising since these *TRBV*s are assumed to perform no functions. However, the functional *TRBV*s, including those of the old and new groups, are more conserved than the pseudo *TRBV*s but possess higher SNPs densities. A possible explanation for this observation is that the functional *TRBV*s are co-evolving with MHC. The MHC molecule is highly diverse in human populations to ensure a diversity of potential responses to antigens. TCRs only recognize the peptides presented by MHC molecules. The most variable regions of MHC are the ones that bind to the antigen. However the regions that bind to the TCR would be more conserved. The CDR1 and CDR2 regions of *TRBV* directly interact with these regions of the MHC molecule. For example, The identity of the polymorphic position 77 in the DRβ chain of class II MHC protein HLA-DR4 co-varies with TCR genes as shown in a previous studies [35]. Thus, it is highly possible that *TRBV* polymorphisms are evolving along with the highly diverse MHC due to the close relationship between these two gene families. A recent study had demonstrated the human TCR-MHC coevolution after divergence from rodents, which suggested a mechanism for ensuring that any V-J gene combination can be selected by a single MHC [7]. It's also reported that there exist strong trans associations between variation in the MHC locus and TCR V gene usage [35]. These findings agree with our conclusion that evolution of TCR is restricted by MHC evolution. The co-evolution might be one of the mechanisms that drive to increase the TCR repertoire diversities in humans [7].

The evolution of V genes has been studied previously based on TCR sequence divergences [11,28,29]. In addition, the co-evolution of TCR and MHC has been proposed in several studies [3,7,35]. However, it remains debated whether TCR-MHC restriction is intrinsic as a consequence of species evolution or is merely a result of T cell thymus selection during T cell maturation [3]. Each possibility has been supported by a number of studies [6,15,30,35]. In our study, the high SNP densities in functional *TRBV*s suggest that MHC-TCR restriction may explain the co-evolution of these gene families, in agreement with the speculation in a previous study that TCR-MHC co-evolution occurs in vertebrate evolution [28]. The usages of new *TRBV*s were no fewer than those of old *TRBV*s, suggesting that the new *TRBV*s had already been integrated into the immune system during human evolution. In addition, the new *TRBV*s possess more SNPs than do the old ones, suggesting that the former may be co-evolving more tightly with MHC genes. These speculations require verification via large-scale genomic analyses in the future, by which highly diverse HLA types were investigated, the corresponding *TRBV* usages in population were identified, and the correlations between these two molecules would be in focus.

In contrast to our expectation, evolutionary rate (Ka/Ks) was not correlated with the expression/usage levels of *TRBV*s in donor peripheral blood. However, the variation in *TRBV* usage (as indicated by Shannon entropy, Gini coefficient and relative standard deviation) was correlated with *TRBV* SNP density, which means that *TRBV*s possessing higher numbers of SNPs would be more stably expressed/used in the peripheral blood among healthy persons. This observation might also be explained by TCR-MHC co-evolution. *TRBV*s with high polymorphisms co-evolve strongly with the MHC, leading to stable usage. *TRBV*s with low polymorphisms may able to fit in the MHC in some portions of

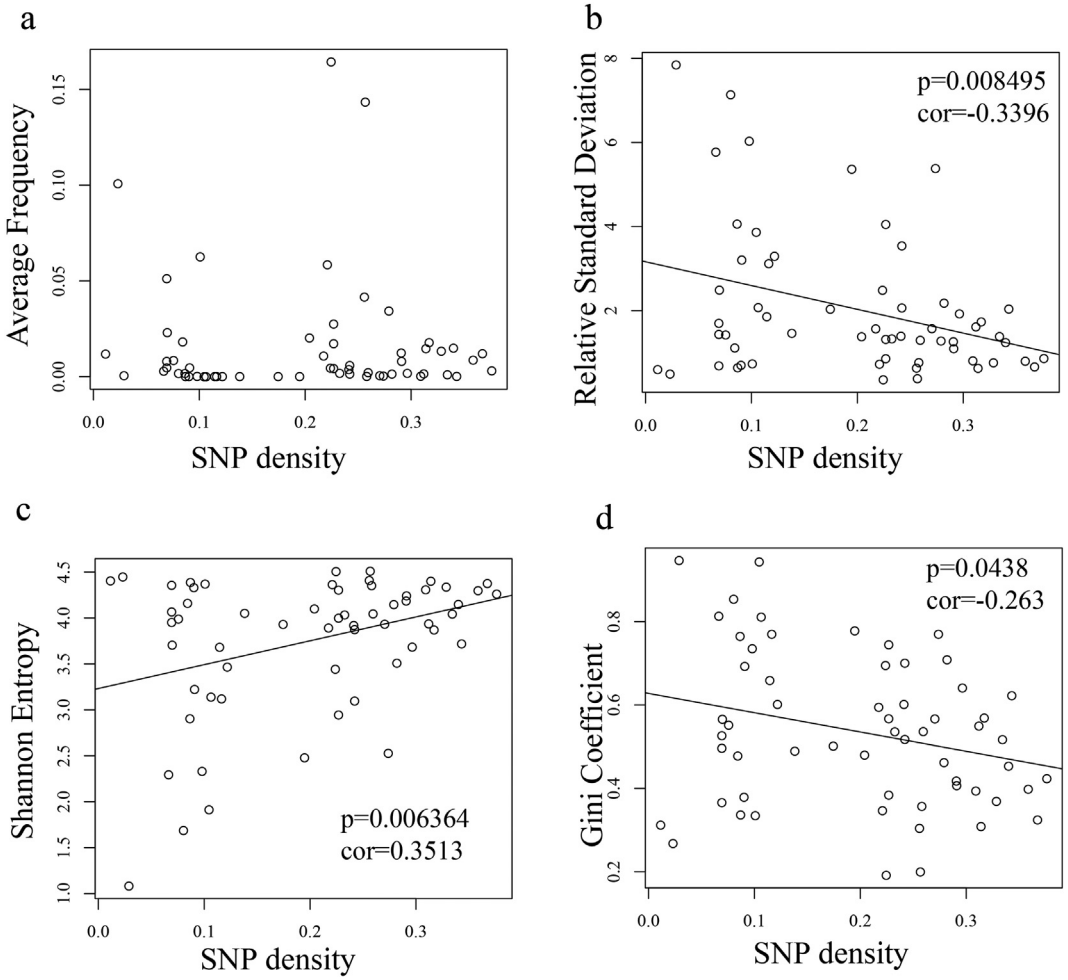

Fig. 5. Correlations between SNP density and *TRBV* distribution indexes in 97 healthy donors. The p-values (p) and correlations (cor) from Pearson correlation tests are presented. a) The correlation between SNP density and average *TRBV* frequency in peripheral blood of a healthy population. No correlation was found. b) The correlation between human *TRBV* SNP density and relative standard deviation of human *TRBV* frequency in peripheral blood of a healthy population. c) The correlation between human *TRBV* SNP density and Shannon entropy of human *TRBV* frequency in peripheral blood of a healthy population. d) The correlation between human *TRBV* SNP density and Gini coefficient of human *TRBV* frequency in peripheral blood of a healthy population.

population but be selected against by the thymus in other portions, leading to variation in usage. This interpretation can explain how MHC restrictions form during evolution.

In this study, all *TRBV* gene (n = 61) sequences were retrieved from the reference *TRB* sequence (GeneID: 6957), which can be completely mapped to the human genome (GRCh38). However, this version of the *TRB* sequence is missing 5 *TRBV*s relative to the *TRBV* genes (n = 66) in IMGT, and the lengths of the two *TRB* sequences are not identical. This indicates that the human genome (GRCh38) may have some flaws in the *TRB* region. More accurate results can be achieved if a more precise genome sequence is used. Another limitation of this study is that all of the donors in this study were Han Chinese from a population in South China. Although *TRBV* usages are not expected to differ among populations from different geographical regions, the conclusions can be strengthened if donors from all over the world are recruited. Finally, our proposal that the SNP levels of *TRBV*s are related to TCR-MHC co-evolution can be evaluated by large-scale genomic association analyses in the future.

5. Conclusions

T cell receptors have been a research hotspot for a long time. However, many of the mechanisms associated with T cell receptors remain unclear, and the co-evolution of TCR and MHC is poorly understood. In this study, we compared human *TRBV*s and their homologous sequences in three reference species, grouped the *TRBV*s into 3 groups based on their features and analyzed the variations of TRBVs in the peripheral blood of 97 donors. The results showed that there were great differences in TRBVs between primates and non-primates. The evolution forces were unequally distributed on the full length of TRB locus, because 5′-half of the TRB locus were undergoing more duplications than the 3′-half. The SNP levels show significant difference among the pseudo, old and new *TRBV*s. Also in population, different expression patterns exist among these three groups of *TRBV*s. Meanwhile, the variations of TRBVs usages exhibit significant correlations with the SNP levels in population. In conclusion, the evolution of *TRBV*s may be influenced by TCR-MHC co-evolution and that the usage of TRBVs in the peripheral blood of humans from a healthy population may be influenced by this co-evolution. This study provides new insights into the evolution of the adaptive immune system.

Disclosure of potential conflicts of interest

The authors declare no potential conflict of interest.

Author contributions

X.M. designed the study, performed the data analyses and wrote the

manuscript. X.C. and J.C. performed the experiments. Y.J. performed the data annotations. Y.P. and C.L. collected the blood samples. K.L. helped revise the manuscript. W.L and F.L. supervised the study.

Funding

This work was funded by Grants from National Natural Science Foundation of China (81402255), Guangdong Province Natural Science Funds for Distinguished Young Scholar (2016A030306050), Science and Technology Innovation Platform in Foshan city (2015AG10002), "Guangdong Te Zhi program" Youth Science and Technology talent of project (2015TQ01R462), Guangdong Province Science and Technology Planning Project (2017A020215185), and Guangdong Province Science and Technology Planning Project (2016A020215005).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.humimm.2018.12.007.

References

[1] R. Antonacci, M. Bellini, A. Pala, M. Mineccia, M.S. Hassanane, S. Ciccarese, S. Massari, The occurrence of three D-J-C clusters within the dromedary TRB locus highlights a shared evolution in Tylopoda, Ruminantia and Suina, Devel. Comp. Immunol. 76 (2017) 105–119.
[2] M. Attaf, E. Huseby, A.K. Sewell, alphabeta T cell receptors as predictors of health and disease, Cellular Mol. Immunol. 12 (2015) 391–399.
[3] B.M. Baker, B.D. Evavold, MHC bias by T cell receptors: genetic evidence for MHC and TCR coevolution, Trends Immunol. 38 (2017) 2–4.
[4] B.D. Bitarello, S. Francisco Rdos, D. Meyer, Heterogeneity of dN/dS Ratios at the Classical HLA Class I Genes over Divergence Time and Across the Allelic Phylogeny, Journal of molecular evolution 82 (2016) 38–50.
[5] S.J. Blevins, B.G. Pierce, N.K. Singh, T.P. Riley, Y. Wang, T.T. Spear, M.I. Nishimura, Z. Weng, B.M. Baker, How structural adaptability exists alongside HLA-A2 bias in the human alphabeta TCR repertoire, Proc. Natl. Acad. Sci. U.SA. 113 (2016) E1276–1285.
[6] S. Bowen, P. Sun, F. Livak, S. Sharrow, R.J. Hodes, A novel T cell subset with trans-rearranged Vgamma-Cbeta TCRs shows Vbeta expression is dispensable for lineage choice and MHC restriction, J. Immunol. 192 (2014) 169–177.
[7] X. Chen, L. Poncette, T. Blankenstein, Human TCR-MHC coevolution after divergence from mice includes increased nontemplate-encoded CDR3 diversity, J. Exp. Med. 214 (2017) 3417–3433.
[8] E.T. Clambey, B. Davenport, J.W. Kappler, P. Marrack, D. Homann, Molecules in medicine mini review: the alphabeta T cell receptor, J. Mol. Med. 92 (2014) 735–741.
[9] N. Degauque, S. Brouard, J.P. Soulillou, Cross-reactivity of TCR repertoire: current concepts, challenges, and implication for allotransplantation, Front. Immunol. 7 (2016) 89.
[10] R.O. Emerson, W.S. DeWitt, M. Vignali, J. Gravley, J.K. Hu, E.J. Osborne, C. Desmarais, M. Klinger, C.S. Carlson, J.A. Hansen, M. Rieder, H.S. Robins, Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire, Nat. Genet. 49 (2017) 659–665.
[11] M.F. Flajnik, M. Kasahara, Origin and evolution of the adaptive immune system: genetic events and selective pressures, Nat. Rev. Genet. 11 (2010) 47–59.
[12] J. Glanville, H. Huang, A. Nau, O. Hatton, L.E. Wagar, F. Rubelt, X. Ji, A. Han, S.M. Krams, C. Pettus, N. Haas, C.S.L. Arlehamn, A. Sette, S.D. Boyd, T.J. Scriba, O.M. Martinez, M.M. Davis, Identifying specificity groups in the T cell receptor repertoire, Nature 547 (2017) 94–98.
[13] J.J. Goronzy, Q. Qi, R.A. Olshen, C.M. Weyand, High-throughput sequencing insights into T-cell receptor repertoire diversity in aging, Genome Med. 7 (2015) 117.
[14] E.E. Harris, D. Meyer, The molecular signature of selection underlying human adaptations, Am. J. Phys. Anthropol. Suppl. 43 (2006) 89–130.
[15] S.J. Holland, I. Bartok, M. Attaf, R. Genolet, I.F. Luescher, E. Kotsiou, A. Richard, E. Wang, M. White, D.J. Coe, J.G. Chai, C. Ferreira, J. Dyson, The T-cell receptor is not hardwired to engage MHC ligands, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) E3111–3118.
[16] L.D. Hurst, N.G. Smith, Do essential genes evolve slowly? Current biology: CB 9 (1999) 747–750.
[17] Y. Ji, A.J. Little, J.K. Banerjee, B. Hao, E.M. Oltz, M.S. Krangel, D.G. Schatz, Promoters, enhancers, and transcription target RAG1 binding during V(D)J recombination, J. Exp. Med. 207 (2010) 2809–2816.
[18] W.J. Kent, BLAT–the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[19] J. Klein, Y. Satta, C. O'HUigin, N. Takahata, The molecular descent of the major histocompatibility complex, Annu. Rev. Immunol. 11 (1993) 269–295.
[20] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (2018) 1547–1549.
[21] H. Liang, W.H. Li, Lowly expressed human microRNA genes evolve rapidly, Mol. Biol. Evol. 26 (2009) 1195–1198.
[22] J. Liu, Y. Zhang, X. Lei, Z. Zhang, Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective, Genome Biol. 9 (2008) R69.
[23] M. Long, E. Betran, K. Thornton, W. Wang, The origin of new genes: glimpses from the young and old, Nat. Rev. Genet. 4 (2003) 865–875.
[24] W. Luo, W.T. He, Q. Wen, S. Chen, J. Wu, X.P. Chen, L. Ma, Changes of TCR repertoire diversity in colorectal cancer after Erbitux (cetuximab) in combination with chemotherapy, Am. J. Cancer Res. 4 (2014) 924–933.
[25] C.A. Moncada, E. Guerrero, P. Cardenas, C.F. Suarez, M.E. Patarroyo, M.A. Patarroyo, The T-cell receptor in primates: identifying and sequencing new owl monkey TRBV gene sub-groups, Immunogenetics 57 (2005) 42–52.
[26] R. Nielsen, Z. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, Genetics 148 (1998) 929–936.
[27] C. Notredame, D.G. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate multiple sequence alignment, J. Mol. Biol. 302 (2000) 205–217.
[28] D.N. Olivieri, S. Gambon-Cerda, F. Gambon-Deza, Evolution of V genes from the TRV loci of mammals, Immunogenetics 67 (2015) 371–384.
[29] Z.E. Parra, M. Lillie, R.D. Miller, A model for the evolution of the mammalian t-cell receptor alpha/delta and mu loci based on evidence from the duckbill Platypus, Mol. Biol. Evol. 29 (2012) 3205–3214.
[30] H.L. Parrish, N.R. Deshpande, J. Vasic, M.S. Kuhns, Functional evidence for TCR-intrinsic specificity for MHCII, Proc. Natl. Acad. Sci. U.S.A. 113 (2016) 3000–3005.
[31] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite, Trends Genet.: TIG 16 (2000) 276–277.
[32] M.G. Rudolph, R.L. Stanfield, I.A. Wilson, How TCRs bind MHCs, peptides, and coreceptors, Annu. Rev. Immunol. 24 (2006) 419–466.
[33] D.G. Schatz, Y. Ji, Recombination centres and the orchestration of V(D)J recombination, Nat. Rev. Immunol. 11 (2011) 251–263.
[34] S.F. Schluter, R.M. Bernstein, H. Bernstein, J.J. Marchalonis, 'Big Bang' emergence of the combinatorial immune system, Dev. Comp. Immunol. 23 (1999) 107–111.
[35] E. Sharon, L.V. Sibener, A. Battle, H.B. Fraser, K.C. Garcia, J.K. Pritchard, Genetic variation in MHC proteins is associated with T cell receptor expression biases, Nature Genet. 48 (2016) 995–1002.
[36] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.
[37] B. Shi, L. Ma, X. He, P. Wu, P. Wang, X. Wang, R. Ma, X. Yao, Compositional characteristics of human peripheral TRBV pseudogene rearrangements, Sci. Rep. 8 (2018) 5926.
[38] W. Vecino, C. Daubenberger, R. Rodriguez, A. Moreno, M. Patarroyo, G. Pluschke, Sequence and diversity of T-cell receptor beta-chain V and J genes of the owl monkey Aotus nancymaae, Immunogenetics 49 (1999) 792–799.
[39] A.C. Villani, S. Sarkizova, N. Hacohen, Systems immunology: learning the rules of the immune system, Annu. Rev. Immunol. 36 (2018) 813–842..
[40] X. Wang, P. Wang, R. Wang, C. Wang, J. Bai, C. Ke, D. Yu, K. Li, Y. Ma, H. Han, Y. Zhao, X. Zhou, L. Ren, Analysis of TCRbeta and TCRgamma genes in Chinese alligator provides insights into the evolution of TCR genes in jawed vertebrates, Dev. Comp. Immunol. 85 (2018) 31–43.
[41] Y.I. Wolf, P.S. Novichkov, G.P. Karev, E.V. Koonin, D.J. Lipman, The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 7273–7280.
[42] L.Y. Yampolsky, M.A. Bouzinier, Faster evolving Drosophila paralogs lose expression rate and ubiquity and accumulate more non-synonymous SNPs, Biology Direct 9 (2014) 2.
[43] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci.: CABIOS 13 (1997) 555–556.
[44] Z. Zhang, J. Xiao, J. Wu, H. Zhang, G. Liu, X. Wang, L. Dai, ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments, Biochem. Biophys. Res. Commun. 419 (2012) 779–781.