ORIGINAL RESEARCH
published: 30 August 2019
doi: 10.3389/fimmu.2019.02064

Germline-Encoded TCR-MHC Contacts Promote TCR V Gene Bias in Umbilical Cord Blood T Cell Repertoire

Kai Gao [1,2,3†], Lingyan Chen [2†], Yuanwei Zhang [2†], Yi Zhao [2,4], Ziyun Wan [2], Jinghua Wu [2], Liya Lin [2], Yashu Kuang [3], Jinhua Lu [3,5], Xiuqing Zhang [1,2], Lei Tian [2], Xiao Liu [1,2*] and Xiu Qiu [3,5,6*]*

[1] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [2] BGI-Shenzhen, Shenzhen, China, [3] Division of Birth Cohort Study, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China, [4] School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, [5] Department of Women and Children's Health Care, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China, [6] Department of Obstetrics and Gynecology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
*Thomas Herrmann,
Julius Maximilian University of Würzburg, Germany*

Reviewed by:
*Philip Bradley,
Fred Hutchinson Cancer Research Center, United States
Philip Anton Van Der Merwe,
University of Oxford, United Kingdom
Tobias L. Lenz,
Max-Planck-Institut für Evolutionsbiologie, Germany*

***Correspondence:**
*Lei Tian
tianlei1@genomics.cn
Xiao Liu
liuxiao@genomics.cn
Xiu Qiu
qxiu0161@163.com;
xiu.qiu@bigcs.org*

† These authors have contributed equally to this work

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

Received: 20 May 2019
Accepted: 15 August 2019
Published: 30 August 2019

Citation:
Gao K, Chen L, Zhang Y, Zhao Y, Wan Z, Wu J, Lin L, Kuang Y, Lu J, Zhang X, Tian L, Liu X and Qiu X (2019) Germline-Encoded TCR-MHC Contacts Promote TCR V Gene Bias in Umbilical Cord Blood T Cell Repertoire. Front. Immunol. 10:2064. doi: 10.3389/fimmu.2019.02064

T cells recognize antigens as peptides bound to major histocompatibility complex (MHC) proteins through T cell receptors (TCRs) on their surface. To recognize a wide range of pathogens, each individual possesses a substantial number of TCRs with an extremely high degree of variability. It remains controversial whether germline-encoded TCR repertoire is shaped by MHC polymorphism and, if so, what is the preference between MHC genetic variants and TCR V gene compatibility. To investigate the "net" genetic association between MHC variations and TRBV genes, we applied quantitative trait locus (QTL) mapping to test the associations between MHC polymorphism and TCR β chain V (TRBV) genes usage using umbilical cord blood (UCB) samples of 201 Chinese newborns. We found TRBV gene and MHC loci that are predisposed to interact with one another differ from previous conclusions. The majority of MHC amino acid residues associated with the TRBV gene usage show spatial proximities in known structures of TCR-pMHC complexes. These results show for the first time that MHC variants bias TRBV gene usage in UCB of Chinese ancestry and indicate that germline-encoded contacts influence TCR-MHC interactions in intact T cell repertoires.

Keywords: TRBV gene usage, MHC genetic variations, quantitative trait locus mapping, umbilical cord blood, TCR-MHC co-evolution

INTRODUCTION

T cell immune surveillance is critical for the health of all jawed vertebrates. Most T cells express αβ TCRs and recognize peptides derived from digested proteins when presented at the cell surface in MHC molecules. How αβ TCR interacts with peptide-MHC (pMHC) has been a particularly attractive field as it may contribute to developing strategies for manipulating T cell responses in many diseases including immunodeficiencies, tumor, autoimmune, and allergic diseases.

Looking at the structures, TCR-pMHC interaction is a delicate process. TCR exists in heterodimers and the binding site of each TCR chain can be divided into three complementarity-determining regions (CDRs), called CDR1, 2, and 3. The most variable region CDR3 is formed by somatic recombination of the variable (V), joining (J), and in β chains diversity (D) genes (1–3). Less variable regions CDR1 and CDR2 loop sequences are constant for each type of chain, and are therefore referred to as "germline-derived." Unlike CDR3, CDR1, and CDR2 regions are encoded by only TCR V genes. As such, the T cell repertoire of each individual may have a potential number of 10^{18} TCRs (4). Human MHC genes are also known as human leukocyte antigen (HLA) genes. HLA genes are extremely polymorphic with more than 12,000 known alleles and MHC haplotypes are highly variable in different ethnic groups (5, 6). Each individual inherits one set of MHC genes from a parent including classical MHC class I loci (HLA-A, -B, and -C) and classical MHC class II loci (HLA-DP, -DQ, and -DR). In the dozens of structures of TCR-pMHC complexes that have been solved, CDR1 and CDR2 loops are shown to in contact with the conserved α-helical residues of the MHC molecules, and the highly variable CDR3 loops primarily interact with the peptide (7–9).

Recently, biased αβTCR repertoires and TCR "signatures" raised against specific antigens have been observed in various diseases, especially in virus infections, autoimmune disorders, and tumor (10). Little is known about whether, and if so how MHC genotype influences the composition of TCR repertoire. Most recently, a genetic study done by Pritchard laboratory showed that MHC gene is the most influential gene of TCR V gene usage using RNA-seq data from a large cohort of European adults (11). However, there are also other pieces of evidence suggesting that germline-encoded TCR sequences are not the primary sites for TCR-pMHC interaction. Structural data show that the binding energetics can focus on the CDR3 contacts with peptide and/or the MHC alpha helices in some complexes (12, 13). In addition, there has been a long debate over whether TCRs have been selected evolutionarily to react with MHC (7, 9, 14). As TCR repertoire change greatly based on antigen experience, exhaustion and peripheral T cell proliferation, especially age-associated changes in primary and secondary lymphoid organs, it would be interesting to see how MHC molecules affect TCR repertoire with confounding factors in control.

Therefore, we investigated whether MHC genetic variations reshape the intact TCR repertoires using umbilical cord blood (UCB) to exclude any influential factors that have yet to be introduced. Most UCB T cells are naïve cells occupied by dominant both naïve $CD4^+$ and naïve $CD8^+$ T cells. Specifically, 87.6 ± 5.2% UCB $CD4^+$ T cells co-expressed CD45RA naive antigen and the percentage is 93.5 ± 7.8% for UCB $CD8^+$ T cells (15). And among all age groups, UCB T cells have the highest TCR diversity, the largest number of pathogen-specific clonotypes with the smallest average clonotype size (16) which mean the least probability of biased clonal amplification caused by environmental antigens. These features mentioned above indicate the naive immunophenotype of UCB T cell repertoire. We obtained UCB samples from 201 Chinese newborns, freshly isolated the buffy coat and extracted DNAs. We then sequenced the DNA samples for TCR β chain V genes and the MHC region and performed QTL mapping to test the associations between TRBV gene usage and MHC variations at three levels of allele, nucleotide and amino acid. Our results identified TRBV gene and MHC loci that are predisposed to interact with one another. Contacts between MHC and TRBV molecules promote a genotype shift in TRBV gene usage.

MATERIALS AND METHODS

Umbilical Cord Blood Collection and DNA Isolation

Six milliliter umbilical cord blood was collected from each of 201 healthy newborns after delivery at Guangzhou Women and Children's Medical Center (Guangzhou city, Guangdong, China). The sample collection was performed in accordance with the ethical standards of the Ethics Committee of Guangzhou Women and Children's Medical Center (GWCMC) and written informed consent was obtained from all participating pregnant women. The buffy coat of cord blood was freshly isolated with density gradient centrifugation and DNA was extracted using the HiPure Blood DNA Mini Kit (Magen) according to the manufacturer's protocol. DNA concentration and integrity were measured by Qubit 3.0 fluorometer (Life Technologies, Paisley, UK) and agarose gel (Agilent) electrophoresis.

Immune Repertoire Library Preparation and Sequencing

The 1.2 μg DNA sample was partitioned to construct a library for TCR β (TRB) chain and immune globulin heavy (IGH) chain sequencing using a two-step-PCR method. Firstly, buffy coat DNA was subjected to Multiplex PCR (MPCR) using published primers and cycling conditions to enrich rearranged complementarity determining region 3 (CDR3) of the variable regions of TRB/IGH chain (17, 18). The second PCR introduced the sequencing primers, Illumina primers P5 and P7, into the first PCR products. IGH and TRB genes were sequenced on the HiSeq 4000 (Illumina, La Jolla, CA) with the standard paired-end 150 and paired-end 100 protocol, respectively. Base calling was performed according to the manufacturer's instruction.

MHC Region Capture, Library Construction, and Sequencing

MHC region capture sequencing was performed using the similar protocol reported before (19) with the same design of MHC capture probes named as 110729_HG19_MHC_L2R_D03_EZ, whose product information is provided on the Roche NimbleGen website (https://sequencing.roche.com/en/technology-research/research/immunogenetics.html). This probe set was designed to template upon the 8 annotated human MHC haplotypes and to encompass the 5 megabases (Mb) of the extended MHC region. The evaluation showed that ∼97% of the MHC region, and over 99% of the genes in MHC region, are covered; 98% of genotypes called by this capture sequencing prove consistent with established HapMap genotypes (19). And the sequencing adaptors are modified to fit BGISEQ-500 sequencer

(BGI-Shenzhen, Shenzhen, China). In brief, the 1 μg shotgun library was hybridized to the capture probes following the manufacturer's protocols (Roche NimbleGen) and the captured target fragments were amplified using AccuPrime® Pfx DNA Polymerase (Invitrogen). The PCR products of captured DNA were purified and quantified by Qubit® dsDNA BR Assay Kits (Invitrogen), and then used to construct sequencing library guided by the manufacturer's protocol (BGISEQ-500) (20) including cyclizing and digestion with enzymes and quantified by Qubit® ssDNA Assay Kit (Invitrogen). Finally, libraries were sequenced with standard paired-end 50 reads on the BGISEQ-500 sequencer following the manufacturer's instructions (20).

Immune Repertoire Analysis

TRB and IGH Sequencing data were analyzed using IMonitor (version 1.3.0) developed by our laboratory previously (17). The whole analysis pipeline could be split into the following steps briefly. Firstly, the basic quality control of raw reads was performed including removing low quality reads with average Phred Quality Score <15 and cleaned paired-end reads were merged with their overlapped nucleotides. Secondly, the merged data was aligned to the V, D, J germline genes and alleles that were deposited in the IMGT database (www.imgt.org) by the BLAST (21–23); V, D, and J gene segments were appointed for each sequence according to its alignment result. Thirdly, low abundance sequences with <5 support reads were removed and the left sequences were translated into amino acid sequences. Finally, the TCR or IGH data, such as V-J pairing, V/J usage, CDR3 sequence frequency and CDR3 length distribution, was counted. The average effective data is ∼2.3 M reads for both TRB (PE100) and IHG (PE150). The average percentages of V gene alignment are 77.91 and 88.22% for TRB and IGH, respectively.

The parameters of IMonitor used in our samples were: -ec -k (100 for TRB or 150 for IGH) -jif 80 -vif 80 -v 33 -d and all the parameters have been used or described in other immune repertoire studies (17, 24, 25). To allow data analysis by QTL mapping, undetectable TRBV and IGHV gene were defined as zero. Log$_2$-transformed usage of TRBV (0.01 pseudo-usage was added to avoid zeroes for TRBV) and IGHV (0.00001 pseudo-usage was added to avoid zeroes for IGHV) were displayed by R code with hierarchical clustering of both rows and columns.

Variant Calling in Raw Reads of MHC Region

For each sequenced sample, the mean coverage per bp was 67.57 times with 335.86 Mb effective data, covering 93.19% of the extended MHC region on average. MHC targeted capture sequencing raw data was first quality controlled by filtering out reads with more than 50% bad bases that have quality score $\leq$ 5, then aligned to the human reference genome sequence (UCSC hg19) using BWA (26) (version 0.5.9). BAM files were firstly processed by Picard (version 1.54, http://broadinstitute.github.io/picard) to sort, merge and mark duplications, and then were managed by Genome Analysis Toolkit (27) (GATK, version 3.4) to recalibrate bases, call variants in HaplotypeCaller mode and recalibrate variant quality scores. Only variants labeled as "PASS" by GATK were kept.

Imputation of MHC Alleles

Using variants called by GATK, we imputed four-digit classical HLA alleles for HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1, HLA-DPB1, HLA-DPA1, and HLA-DQA1 with Beagle (28) (version 4.1). The Han-MHC reference panel (29) (total number of individuals 10,689) was used for imputation. In the evaluation, the aggregated mean concordance (r^2 value) was 0.97 for common, 0.93 for low-frequency and 0.81 for rare HLA alleles. Overall, the concordance rate was about 0.93 at four-digit resolution and 0.97 at two-digit resolution when the eight most polymorphic HLA genes were evaluated (29). For each sample and each gene, MHC alleles with the highest genotype probability (estimated by Beagle) were selected. SNPs corresponding to MHC alleles were obtained by aligning exon sequences from the IMGT/HLA database (5) to the human reference genome sequence (UCSC hg19) using BWA (version 0.5.9). Amino acid polymorphisms corresponding to MHC alleles were obtained by annotation of SNPs using ANNOVAR (30) (version 2017 July 16).

QTL Mapping

The genetic variations at single nucleotide, amino acid, and four-digit classical MHC haplotypes were coded as allelic dosage counting on the number of reference allele/s (0, 1, 2). For example, if an SNP is bi-allelic with A and C alleles, we take allele A as the reference allele and count the number of A allele for each individual and code the genotype AA, AC, and CC as 2, 1, and 0. If an SNP is tri-allelic with three possible alleles, then we take each of the possible alleles as reference allele and the others as alternatives, generating three SNP IDs for each tri-allelic SNP. For example, if an SNP has A, C, and T, then we will have three IDs for this SNP: SNP1 will take A as the reference allele and treat the other two alleles as alternatives, therefore, an SNP with genotype of AA is coded as 2, AT and AC are both coded as 1, TT and CC are both coded as 0, and so are for the other two SNP IDs. One tri-allelic SNP is then treated as three SNPs for further analysis.

We used the MatrixEQTL R package (31) for QTL mapping of TRBV and IGHV genes usage with genetic variations in the MHC region. We applied QTL mapping at three levels of genetic variations. SNPs, polymorphic amino acids and four-digit MHC alleles with a MAF (minor allele frequency) < 0.05 were removed. A threshold of 5% FDR was used to control for the testing burdens of multiple variations at each genetic level.

Conditional Analysis of Associations Between TRBV Genes Usage and MHC Amino Acid Variations

The Conditional analysis was performed individually for each TRBV gene usage using forward stepwise linear regression. We established the optimal threshold to perform the stepwise conditional regression on amino acid variations with FDR $<$ 0.05 from the QTL mapping and we considered an expanded model that included the dosage variable for target amino acid variant and the most significant amino acid variant (top amino acid variant) as a covariate. If the target amino acid variations had a conditional $P < 0.05$, we considered it as an independent signal (second amino acid variant). The multiple linear regression model then expanded to include the top amino acid variant and the second amino acid variant as covariates and test the association of the remaining amino acid variants. Regression stopped when the conditional *P*-value of the target amino acid variant was >0.05. We also carried out this type of analysis for the SNPs.

Estimating the Fraction of Variation for Each TRBV Gene Usage That Is Explained by Genetic Variation in the MHC Locus

We fitted a linear regression or a multiple linear regression model (when appropriate) with the target TRBV gene usage as a dependent variable and the significant amino acid variants identified from QTL mapping and conditional analyses as independent variables using *lm* function in R. The proportion of variability explained by the corresponding independent amino acid variant was the adjusted R-square derived from the linear regression model. The statistical significance of the overall model was tested using F-tests. These analyses were also performed at the level of SNPs.

Structure Analysis

Our QTL mapping results of a total of 96 MHC amino acid variations were marked onto protein structures of pMHC complexes. Structures of proteins were downloaded from the RCSB PDB [PDB ID: 2BNQ (32), PDB ID: 4G9F (33), PDB ID: 1EFX (34), PDB ID: 4OZF (35), PDB ID: 1J8H (36)] and were plotted with the UCSF Chimera package (37). We next collected all structures of TCRs bound to pMHCs and all contacting information between corresponding chains of HLA-A, HLA-B, HLA-DRB1, HLA-DQA1, or HLA-DQB1 and any of the TCR β chains or the presented peptides in the IMGT database (5) and calculated the frequency of the presence of contacting amino acid position among the analyzed TCR-pMHC complex. The identified 96 amino acid variants were then directly compared with TCR β and peptide contacting MHC positions.

RESULTS

TRBV Gene Usage Is Influenced by MHC Genetic Variation

To estimate the usage of TRBV genes, we sequenced TCR β chain V genes using DNA collected from UCB of 201 Chinese newborns. Then we calculated the proportion of reads that mapped uniquely to each Vβ gene of all mapped reads. The MHC region of the same individual was sequenced by target capture sequencing and classical four-digit alleles were imputed from SNP data. A schematic overview of the analysis is presented in **Figure 1A.**

We extracted 48 TRBV genes in each of the 201 samples using immune repertoire sequencing data (**Supplementary Figure 1**, **Supplementary Tables 1**, **2**). Undetected TRBV genes are randomly scattered, which is most likely due to individual differences. Unique alleles of 23 HLA-A, 46 HLA-B, 21 HLA-C, 28 DRB1, 16 DPB1, 4 DPA1, 14 DQB1, and 15 DQA1 are present in our cohort (**Supplementary Table 3**). We then performed QTL mapping between TRBV usage and MHC alleles. As a control, B cell Immunol Globulin Heavy chain V (IGHV) genes, which are not expected to interact with MHC, were analyzed in a similar pipeline. Our results showed that the frequencies of TRBV genes, but not the IGHV genes, are significantly associated with MHC alleles (**Figure 1B**, **Supplementary Tables 4**, **5**). We then performed QTL mapping using single nucleotide polymorphisms (SNPs) (**Supplementary Tables 6**, **7**) and amino acid variations (**Supplementary Tables 8**, **9**) corresponding to MHC alleles. Again, only TRBV gene usage is significantly associated with MHC at SNP and amino acid levels (**Figure 1B**). These are the first genetic evidence that TRBV gene usage is significantly influenced by MHC genes in UCB T cell repertoires.

TRBV Genes Differ in Their Association With Different MHC Locus

Next, we aim to find which TRBV genes and MHC variants are most likely to be in strong associations. Our results show that the frequencies of 8.3% (4/48) of TRBV genes can be explained by MHC alleles and the frequencies of 14.6% (7/48) of TRBV genes can be explained by MHC nucleotide and/or amino acid variations (**Figure 2A**). Among the seven TRBV genes influenced by MHC, TRBV13, TRBV7-6, TRBV7-9, TRBV10-3, and TRBV30 are novel; TRBV20-1 and TRBV9 have been described before in European adults (11). In accordance with the previous report, our results indicate that TCR V genes differ in their compatibilities with MHC polymorphism.

Looking at the distribution of significant associations at different MHC loci, 9 MHC alleles, 114 MHC SNPs and 96 MHC amino acid variations are associated with TRBV usage at FDR<0.05 (**Supplementary Tables 4**, **6**, **8**). MHC class I locus HLA-A and MHC class II locus DRB1 each have a larger number of variations associated with TRBV genes than that of any other loci of the same class at both nucleotide and amino acid level (**Figures 2B,C**). If the specificity of TCRs toward MHC molecules depends on only the conservative regions of MHC molecules, then we may expect to find the number of TCR associated MHC molecules to be approximately proportional to the number of MHC variations in each region. Interestingly, HLA-C locus has the minimum number of variations in association with TRBV genes despite the fact that the degree of HLA-C loci diversity is as high as that of HLA-A loci. Although HLA-C shares sequence homology with HLA-A and HLA-B molecules, it is substantially different from other MHC I molecules in many different ways. One of the main feature distinguishing HLA-C is its low expression, that the number of surface HLA-C proteins are estimated to be only ∼10% of that of HLA-A and HLA-B molecules (38–40). Thus, it is highly possible TCR usage bias explained by different MHC loci is also shaped by the intrinsic differences in the abundance of surface MHC proteins, suggesting a strong influence from direct physical contacts.

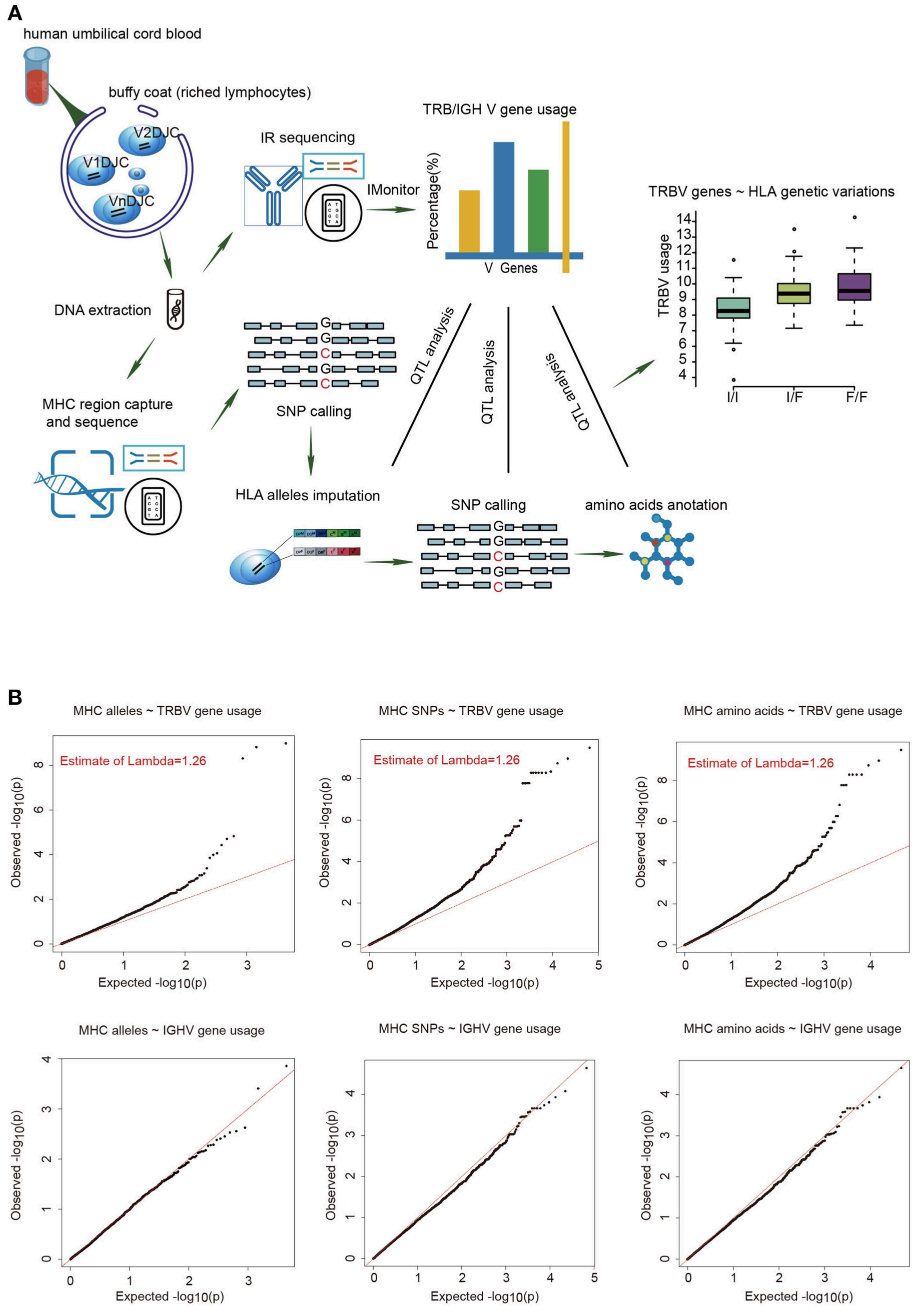

FIGURE 1 | The frequencies of TRBV genes are significantly associated with variations in the MHC locus. **(A)** Schematic overview of the analysis. Usage of TCR β chain V genes was estimated by mapping buffy coat DNA sequencing reads to the human TRBV gene database. MHC alleles were imputed with Beagle. SNPs were called using BWA (version 0.5.9). Amino acid polymorphisms corresponding to MHC alleles were obtained by annotation of SNPs using ANNOVAR. The associations of Vβ usage with nucleotide and amino acid genotypes were tested by QTL mapping using a linear regression model. **(B)** QQ plots for the associations between MHC alleles/SNPs/amino acid variations and the usage of TRBV genes (up) or IGHV genes (down). The red line refers to a normal distribution; each black dot represents one allele/SNP/amino acid variation. Estimated inflation factor lambda is provided if it is statistically significant.

Independent MHC Residues That Bias TRBV Gene Usage

Due to the strong linkage disequilibrium (LD) structure in the MHC region, it is unclear which MHC variant is responsible for a particular association. Therefore, we performed conditional analysis for each TRBV genes to identify their potential independent MHC amino acid variants. For instance, we examined all TRBV13-associated MHC amino acids. Twenty

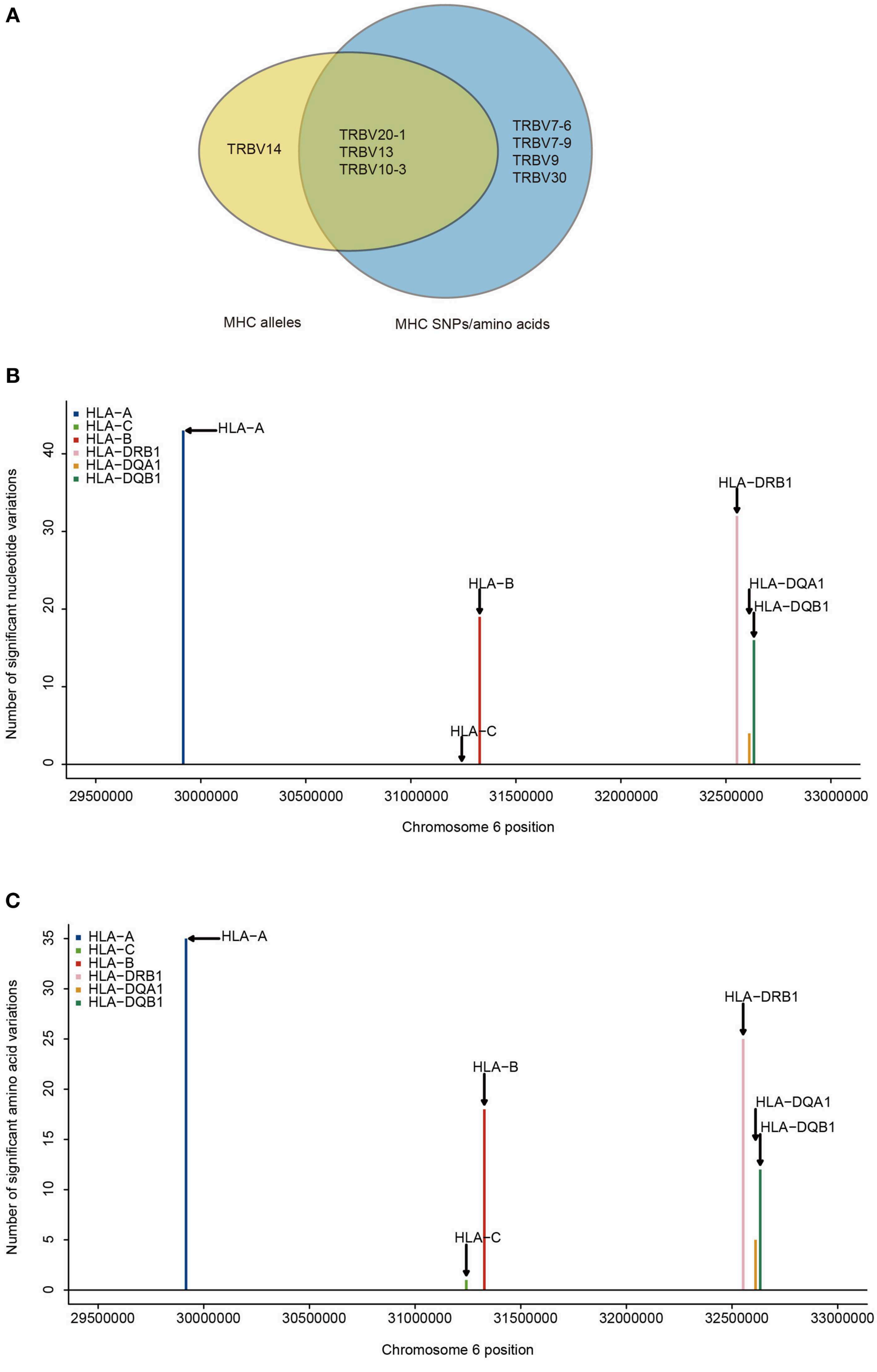

FIGURE 2 | TRBV genes differ in their association with different MHC locus. **(A)** TRBV gene usage explained by MHC alleles (yellow), nucleotide, and amino acid variations of the MHC alleles (blue). The associations of Vβ gene usage with the number of nucleotide variations **(B)**, and with the number of amino acid variations **(C)** (FDR ≤ 0.05). SNPs are binned according to their genomic position.

polymorphic amino acids in HLA-A, five in HLA-B and one in HLA-DRB1 (position 71) showed significant association with TRBV13 (FDR<0.05). Except HLA-A position 97 and HLA-DRB1 position 71, other amino acid positions initially have four potential linkage groups (**Figure 3A**). Our conditional analysis further confirmed that amino acid position 97 of HLA-A have the strongest association with TRBV13, followed by position 71 of HLA-DRB1 (**Figures 3B, C**, conditional $P < 0.05$). After

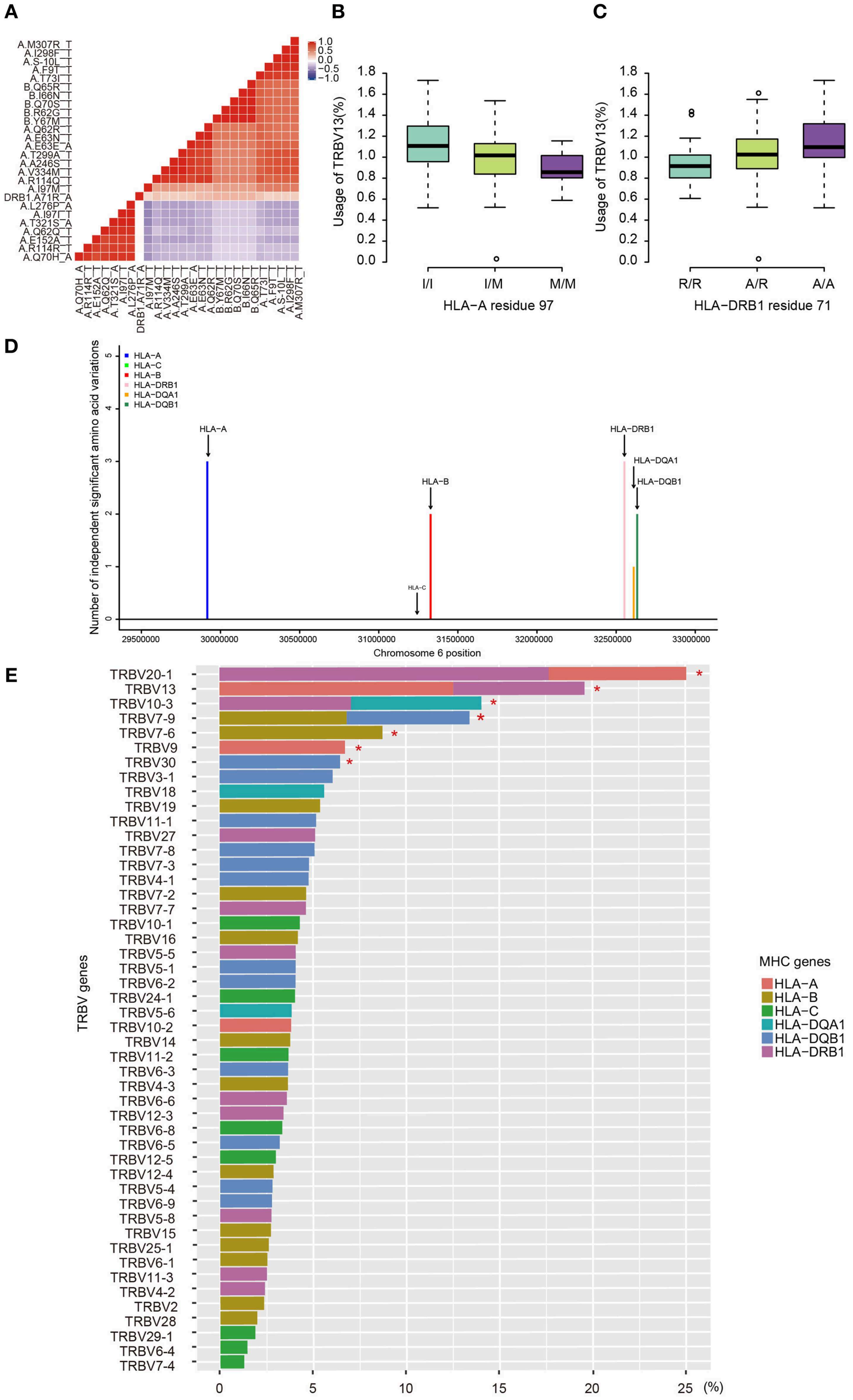

FIGURE 3 | Independent associations between TRBV usage and amino acids variations in MHC alleles. **(A)** The heat map representing a color-coded correlation matrix of all the 26 MHC amino acids that are significantly associated with TRBV13 usage (FDR ≤ 0.05). The frequencies of TRBV13 influence by the two independent amino acid variations at HLA-A residue 97 **(B)** and HLA-DRB1 residue 71 **(C)**. **(D)** The number of the independent Vβ gene associated MHC amino acid variations. **(E)** Variation of TRBV gene usage explained by MHC amino acids variations of the MHC alleles. Values were adjusted R-square derived from the linear regression model. A star indicates that the total proportion of variation explained by the MHC gene components was significant at 5% FDR.

TABLE 1 | Usage variation of TRBV genes explained by independent amino acid variations in the MHC locus.

TRBV gene ID	Top signal	Second signal	% of usage variation explained by top signal	% of usage variation explained by second signal	% of usage variation explained by both signals
TRBV20-1	HLA-DRB1.I67F	HLA-A.D116Y	17.68110065	7.374011699	23.3005267
TRBV13 *	HLA-A.I97M	HLA-DRB1.A71R	12.5520188	7.0425016	17.0733699
TRBV7-6*	HLA-B.V282I	NA	8.74398907	NA	8.74398907
TRBV7-9*	HLA-B.E45K	HLA-DQB1.L75V	6.829646821	6.589435636	12.4010995
TRBV10-3*	HLA-DRB1.D57S	HLA-DQA1.A199T	7.061965751	6.994849504	15.6205873
TRBV9	HLA-A.L276P	NA	6.729255542	NA	6.72925554
TRBV30*	HLA-DQB1.R55P	NA	6.468371558	NA	6.46837156

*****_TRBVs are novel findings from our data._

conditioning for these two top positions, no other associations reach a significant level.

Using this method, 11 independent amino acid variants are found to have the highest possibilities of influencing TRBV gene usage (**Figure 3**, **Supplementary Figure 2**). Three of the 11 independent amino acid variations, HLA-DRB1 residue 67 and 71, HLA-DQB1 residue 55 are associated with type 1 diabetes, multiple sclerosis, hypothyroidism, rheumatoid arthritis, and inflammatory polyarthritis according to PheWAS database (41, 42). We showed the number of independent amino acid variations located in specific MHC genes and the number is comparable between class I and II (**Figure 3D**). Then, we estimated to what extent these MHC variants can influence each TRBV gene usage variation using multiple linear regression. The results showed that 6–23.3% of the proportion of TRBV variability can be explained by MHC amino acid variations, highlighting an important role of MHC in shaping TCR repertoires (**Figure 3E**, **Table 1**).

TRBV Gene Associated MHC Amino Acids Locate in TCR Binding Pockets

MHC residues near the contact interface between TCR and MHC or located in the polymorphic "pockets" of the peptide-binding grooves are most likely to influence the TCR-pMHC (peptide-major histocompatibility complex) interaction (9). To see whether TRBV gene associated MHC residues are adjacent to or have direct contacts with TCRs on the molecular structure, we mapped these MHC residues onto protein structures of pMHC complex downloaded from PDB database (**Figure 4A**). We then collected all structural information of TCR-pMHC complex consisting of 66 HLA-A, 23 HLA-B, 12 HLA-DRB1, 8 HLA-DQA1, or 8 HLA-DQB1 and their paired TCR and peptides (**Supplementary Table 10**). The residues with a high possibility of influencing TRBV gene usage tend to be either physically near or in direct contact with the TCR in structures (**Figure 4B**). For instance, HLA-DRB1 residue at position 67 predicted in our model with the highest association with TRBV20-1 showed contact with TCR in 42% analyzed complex and with peptide in 75% analyzed complex. These results suggest that MHC amino acid residues influence TCR-pMHC interactions via direct physical contacts.

DISCUSSION

Our results show that MHC polymorphism plays an important role in shaping UCB TCR repertoires. 14.6% (7/48) of TRBV genes are significantly associated with nucleotide and/or amino acid variations of the MHC molecules. Among these TRBV associated MHC loci, we are able to pinpoint 11 independent influential MHC amino acid residues, the majority of which are located in HLA-A and HLA-DRB1 loci. The structural analysis confirmed that the majority of TRBV associated MHC residues are positioned at the TCR-pMHC contact interfaces of known protein complexes, which indicates that MHC molecules have a higher probability of influencing TRBV gene frequencies through physical contacts. In summary, we conclude that MHC variations sculpt UCB TRBV gene repertoires by favoring more compatible TCR-MHC pairs in thymic selection.

Our results shed light on the long debate about the basis of TCR specificity for MHC molecules. In 1971, Jerne proposed that TCR and MHC genes coevolve to have inherent predisposition to interact with one another (43). Jerne's idea was further extended that TCRs' biases for MHC are dictated by conserved CDR1 and CDR2 loops encoded by TCR V genes (44), which was later validated in many known TCR-pMHC complexes (7, 45–48). In contradiction to this theory, a human TCR complex with an MHC class II molecule demonstrated no dependent of CDR1 and CDR2 for MHC recognition (12). In this complex, CDR1 and CDR2 have a few contacts with any kind of MHC, and CDR3s have extensive contacts with the peptide and the α helices (12). It is currently impossible, or probably unnecessary, to thoroughly exam all possible TCR-pMHC structures to conceive the rules for TCR-pMHC interactions. Our results provided genetic evidence that intrinsic TCR-MHC bias exists in the UCB samples of a relatively large cohort of newborns. A recent study consistent with this idea comes from Pritchard laboratory, which utilized RNA-seq data from adult peripheral blood (11). In summary, our genetic results are supportive of the inherited reactivity of TCR-MHC molecules.

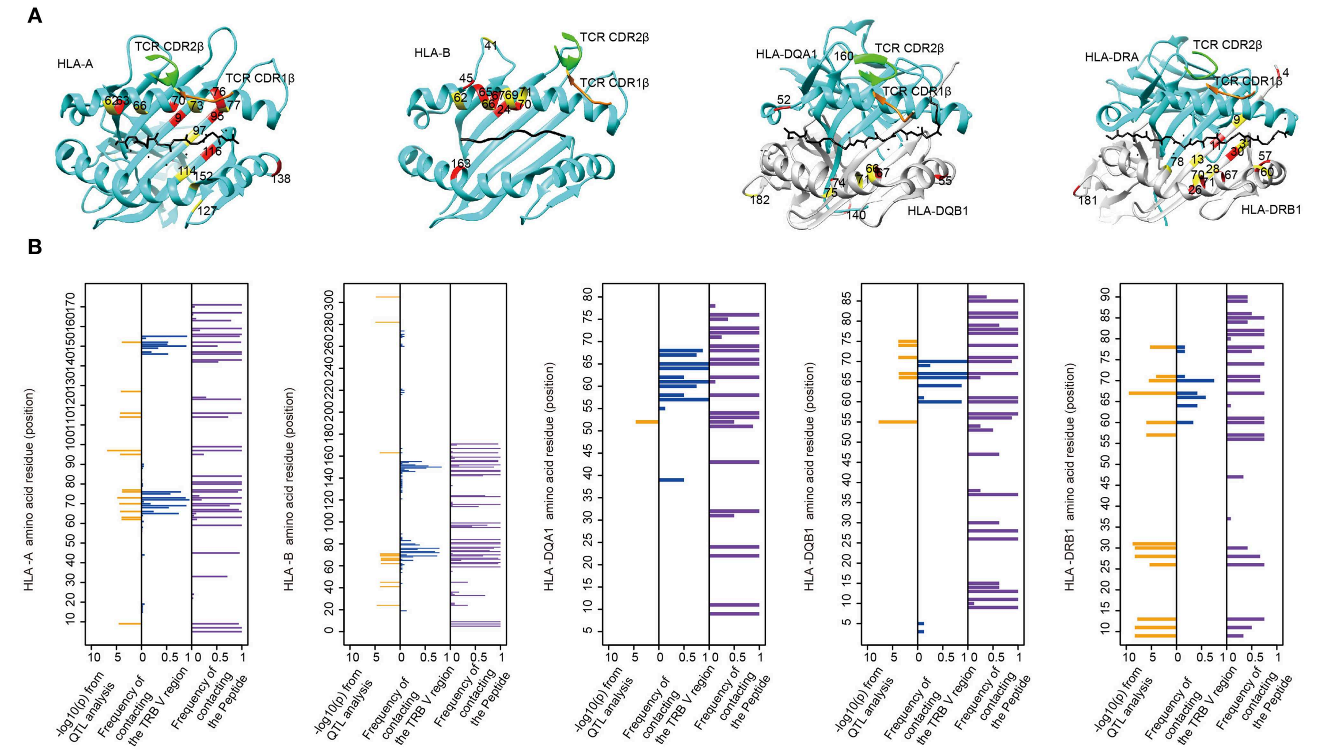

FIGURE 4 | The associated MHC residues tend to be at the TCR-pMHC interface. **(A)** Mapping MHC amino acids biasing the TRBV genes usage (red and yellow) onto a structure of MHC genes (cyan) with CDR1 β chain (orange) and CDR2 β chain (green) (PDB ID from left to right: 2BNQ, 4G9F, 4OZF, and 1J8H). **(B)** Comparison of the Log_{10}-transformed *P*-values for the QTL analysis (left in each panel) and the frequency with which these amino acids physically contact the TCR β chain V region (middle in each panel) or the Peptide (right in each panel) in solved complexes (from left to right: 66 HLA-A complexes, 23 HLA-B complexes, 8 HLA-DQA1 complexes, 8 HLA-DQB1 complexes, and 12 HLA-DRB1 complexes).

Previous studies on the TCR bias for MHC molecules have utilized adult peripheral blood samples of mouse and human. When using adult samples, precautions need to be taken into consideration to exclude confounding factors that may substantially shape T cell repertoire. Two such factors are the thymic involution during aging and prevalent pathogen infections. To minimize any known and unknown confounding factors, we utilized the newborns' UCB samples that allow us to assess the "net" genetic association between TCR and MHC. The results of our UCB data differ from previous data: HLA-A and HLA-B, instead of HLA-C genes, are found to be the most influential MHC class I loci to TRBV gene usage. Although the conservative regions of HLA-C proteins are similar to that of HLA-A and HLA-B proteins, HLA-C is substantially different from other MHC I molecules, especially the surface expression of HLA-C proteins is much lower than those of the HLA-A and HLA-B proteins (49). In addition, only a minority of CD8 T cells are restricted by HLA-C, and they have recently been found to be critical in response to chronic infections such as Epstein-Barr virus and HIV infection (50). Therefore, the associations between HLA-C and TCRs found in the circulation of adults may be biologically meaningful, but may also reflect the prevalence of HLA-C restricted T cell responses.

It is interesting that several MHC loci stand out as more influential than other loci in shaping TRBV genes. There may be important biological meanings in the selective involvement of certain MHC loci in thymic education. The conventional TCR-MHC system is not the only function of the MHC genes or necessarily the original function from an evolution point of view. For instance, many non-classical MHC class I molecules, as well as HLA-C, also function as a ligand for killer immunoglobulin receptors (KIRs) to regulate nature killer (NK) cell activities. Crystal structures of the KIR2DL2-HLA-Cw3 and KIR2DL1-HLA-Cw4 complexes revealed the precise contacting site of HLA-C with KIR molecules, which is predicted to result in KIR competing with TCR for HLA-C interaction (40). Similarly, in the risk predictions of MHC mismatching transplantations, HLA-A, -B, -C, and -DRB1 mismatches are associated with higher risks of graft-vs.-host response than those of other classical MHC loci (51). Together, these results suggest a modification to the applicable MHC loci of the coevolution theory.

With regard to the TRBV genes, we found seven MHC biased TRBV genes among which, TRBV13, TRBV7-6, TRBV7-9, TRBV10-3, and TRBV30 are first reported, and TRBV20-1 and TRBV9 have been described before in European adults (11). Notably, structures that composed of TRBV13 has been extensively studied in mouse. Complexes that include TRBV13-2 show that amino acids in its CDR2 loop react with related sites on the MHCII α1 helix despite various docking angles of the TCR, and TRBV13-1 CDR1 and CDR2 loops have more than one docking site on α1 helix and shifts according to docking positions (52). A dynamic interplay between TCR and MHC molecules has gathered more and more evidence. It is possible that we could eventually construct a list of conserved interactions between TCR and MHC with genetic studies, even though, the two molecules may not interact in a conventional way. One of our limitations is that although we identified the significant associated MHC residues with TRBV gene and explored the structural mechanism, we did not calculate the preference for contact with MHC molecules from the perspective of specific amino acids of TCR. It is therefore not structurally validated and compared to our current results from the MHC amino acid perspective. And we think it would be a novel field or an inspiring perspective to be developed further.

In summary, our results showed that TCR Vβ genes are significantly associated with the MHC genotypes in the UCB from a cohort of 201 Chinese newborns. Our structure analysis suggests that MHC amino acid residues associated with the TRBV gene usage are in contact or adjacent with the TCRβ chains in physical structure showing the substantial potential of MHC to influence TCR in protein-protein interaction. Our results confirm and extend our knowledge of TCR-MHC association and contribute to the richness of side-by-side comparisons of population-based studies as a strategy to further understanding the nature of TCR-MHC interactions and its implications in human.

DATA AVAILABILITY

The datasets that support the findings of this study are deposited under supervision and control in the China National Genebank (CNGB, https://db.cngb.org/cnsa/). Data of this project could be accessed after an approval application. Please refer to https://db.cngb.org/, or email: CNGBdb@cngb.org for detailed application guidance. The accession code CNP0000425 should be included in the application.

AUTHOR CONTRIBUTIONS

XQ and XL conceived and provided overall guidance. LT redesigned the analysis. KG, LC, and YuZ carried out the analysis. LL, JW, YK, and JL carried out the experiments. ZW and YiZ performed the data curation. KG and LT wrote the manuscript. XZ supervised the implementation of the project.

FUNDING

This research was supported by the National Natural Science Foundation of China (81673181 and 31700794) and the Shenzhen Municipal Government of China (JCYJ20170817145404433 and JCYJ20170817145428361). The funders had no role in the study design, data collection, and analysis, the decision to publish or the preparation of the manuscript.

ACKNOWLEDGMENTS

We also thank the donors for providing UCB samples, colleagues of China National GeneBank, BGI-Shenzhen, for their help in producing the solid data, and Yafeng Zhu and Xiaowei Jiang for their help in revising the manuscript. This manuscript has been released as a Pre-Print at bioRxiv (https://www.biorxiv.org/content/10.1101/621821v1) (53).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2019.02064/full#supplementary-material

REFERENCES

1. Cobb RM, Oestreich KJ, Osipovich OA, Oltz EM. Accessibility control of V(D)J recombination. *Adv Immunol.* (2006) 91:45–109. doi: 10.1016/S0065-2776(06)91002-5
2. Little AJ, Matthews A, Oettinger M, Roth DB, Schatz DG. Chapter 2 - the mechanism of V(D)J recombination. In: Alt FW, Honjo T, Radbruch A, Reth M, editors. *Molecular Biology of B Cells.* 2nd ed. London: Academic Press (2015). p. 13–34.
3. Liu X, Wu J. History, applications, and challenges of immune repertoire research. *Cell Biol Toxicol.* (2018) 34:441–57. doi: 10.1007/s10565-018-9426-0
4. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature.* (1988) 334:395–402. doi: 10.1038/334395a0
5. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* (2015) 43(Database issue):D423–31. doi: 10.1093/nar/gku1161
6. La Gruta NL, Gras S, Daley SR, Thomas PG, Rossjohn J. Understanding the drivers of MHC restriction of T cell receptors. *Nat Rev Immunol.* (2018) 18:467–78. doi: 10.1038/s41577-018-0007-5
7. Marrack P, Scott-Browne JP, Dai S, Gapin L, Kappler JW. Evolutionarily conserved amino acids that control TCR-MHC interaction. *Annu Rev Immunol.* (2008) 26:171–203. doi: 10.1146/annurev.immunol.26.021607.090421
8. Collins EJ, Riddle DS. TCR-MHC docking orientation: natural selection, or thymic selection? *Immunol Res.* (2008) 41:267–94. doi: 10.1007/s12026-008-8040-2
9. Garcia KC, Adams EJ. How the T cell receptor sees antigen–a structural view. *Cell.* (2005) 122:333–6. doi: 10.1016/j.cell.2005.07.015
10. Wang CY, Yu PF, He XB, Fang YX, Cheng WY, Jing ZZ. alphabeta T-cell receptor bias in disease and therapy (Review). *Int J Oncol.* (2016) 48:2247–56. doi: 10.3892/ijo.2016.3492
11. Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet.* (2016) 48:995–1002. doi: 10.1038/ng.3625
12. Hahn M, Nicholson MJ, Pyrdol J, Wucherpfennig KW. Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune T cell receptor. *Nat Immunol.* (2005) 6:490–6. doi: 10.1038/ni1187
13. Borg NA, Ely LK, Beddoe T, Macdonald WA, Reid HH, Clements CS, et al. The CDR3 regions of an immunodominant T cell receptor dictate the 'energetic

landscape' of peptide-MHC recognition. *Nat Immunol.* (2005) 6:171–80. doi: 10.1038/ni1155

14. Lu J, Van Laethem F, Bhattacharya A, Craveiro M, Saba I, Chu J, et al. Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nat Commun.* (2019) 10:1019. doi: 10.1038/s41467-019-08906-7
15. D'Arena G, Musto P, Cascavilla N, Di Giorgio G, Fusilli S, Zendoli F, et al. Flow cytometric characterization of human umbilical cord blood lymphocytes: immunophenotypic features. *Haematologica.* (1998) 83:197–203.
16. Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, et al. Dynamics of individual T cell repertoires: from cord blood to centenarians. *J Immunol.* (2016) 196:5005–13. doi: 10.4049/jimmunol.1600005
17. Zhang W, Du Y, Su Z, Wang C, Zeng X, Zhang R, et al. IMonitor: a robust pipeline for TCR and BCR repertoire analysis. *Genetics.* (2015) 201:459–72. doi: 10.1534/genetics.115.176735
18. Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, et al. Systematic comparative evaluation of methods for investigating the TCRbeta repertoire. *PLoS ONE.* (2016) 11:e0152464. doi: 10.1371/journal.pone.0152464
19. Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, et al. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. *PLoS ONE.* (2013) 8:e69388. doi: 10.1371/journal.pone.0069388
20. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, et al. A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience.* (2017) 6:1–9. doi: 10.1093/gigascience/gix024
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* (1990) 215:403–10. doi: 10.1016/S0022-2836(05)80360-2
22. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* (2000) 7:203–14. doi: 10.1089/10665270050081478
23. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* (2006) 34(Web Server issue):W6–9. doi: 10.1093/nar/gkl164
24. Wang L, Zhang W, Lin L, Li X, Saksena NK, Wu J, et al. A comprehensive analysis of the T and B lymphocytes repertoire shaped by HIV vaccines. *Front Immunol.* (2018) 9:2194. doi: 10.3389/fimmu.2018.02194
25. Wang T, Wang C, Wu J, He C, Zhang W, Liu J, et al. The different T-cell receptor repertoires in breast cancer tumors, draining lymph nodes, and adjacent tissues. *Cancer Immunol Res.* (2017) 5:148–56. doi: 10.1158/2326-6066.CIR-16-0107
26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [Preprint].* arXiv:1303.3997 (2013).
27. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* (2013) 43:11.10.1–33. doi: 10.1002/0471250953.bi1110s43
28. Browning BL, Browning SR. Genotype imputation with millions of reference samples. *Am J Hum Genet.* (2016) 98:116–26. doi: 10.1016/j.ajhg.2015.11.020
29. Zhou F, Cao H, Zuo X, Zhang T, Zhang X, Liu X, et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat Genet.* (2016) 48:740–6. doi: 10.1038/ng.3576
30. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* (2010) 38:e164. doi: 10.1093/nar/gkq603
31. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* (2012) 28:1353–8. doi: 10.1093/bioinformatics/bts163
32. Chen JL, Stewart-Jones G, Bossi G, Lissin NM, Wooldridge L, Choi EM, et al. Structural and kinetic basis for heightened immunogenicity of T cell vaccines. *J Exp Med.* (2005) 201:1243–55. doi: 10.1084/jem.20042323
33. Ladell K, Hashimoto M, Iglesias MC, Wilmann PG, McLaren JE, Gras S, et al. A molecular basis for the control of preimmune escape variants by HIV-specific CD8^{+} T cells. *Immunity.* (2013) 38:425–36. doi: 10.1016/j.immuni.2012.11.021
34. Boyington JC, Motyka SA, Schuck P, Brooks AG, Sun PD. Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. *Nature.* (2000) 405:537–43. doi: 10.1038/35014520
35. Petersen J, Montserrat V, Mujico JR, Loh KL, Beringer DX, van Lummel M, et al. T-cell receptor recognition of HLA-DQ2-gliadin complexes associated with celiac disease. *Nat Struct Mol Biol.* (2014) 21:480–8. doi: 10.1038/nsmb.2817
36. Hennecke J, Wiley DC. Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. *J Exp Med.* (2002) 195:571–81. doi: 10.1084/jem.20011194
37. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem.* (2004) 25:1605–12. doi: 10.1002/jcc.20084
38. Snary D, Barnstable CJ, Bodmer WF, Crumpton MJ. Molecular structure of human histocompatibility antigens: the HLA-C series. *Eur J Immunol.* (1977) 7:580–5. doi: 10.1002/eji.1830070816
39. Neefjes JJ, Ploegh HL. Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. *Eur J Immunol.* (1988) 18:801–10. doi: 10.1002/eji.1830180522
40. Blais ME, Dong T, Rowland-Jones S. HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? *Immunology.* (2011) 133:1–7. doi: 10.1111/j.1365-2567.2011.03422.x
41. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med.* (2017) 9:eaai8708. doi: 10.1126/scitranslmed.aai8708
42. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol.* (2013) 31:1102–10. doi: 10.1038/nbt.2749
43. Jerne NK. The somatic generation of immune recognition. *Eur J Immunol.* (1971) 1:1–9. doi: 10.1002/eji.1830010102
44. Sim BC, Zerva L, Greene MI, Gascoigne NR. Control of MHC restriction by TCR Valpha CDR1 and CDR2. *Science.* (1996) 273:963–6. doi: 10.1126/science.273.5277.963
45. Parrish HL, Deshpande NR, Vasic J, Kuhns MS. Functional evidence for TCR-intrinsic specificity for MHCII. *Proc Natl Acad Sci USA.* (2016) 113:3000–5. doi: 10.1073/pnas.1518499113
46. Adams JJ, Narayanan S, Birnbaum ME, Sidhu SS, Blevins SJ, Gee MH, et al. Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity. *Nat Immunol.* (2016) 17:87–94. doi: 10.1038/ni.3310
47. Dai S, Huseby ES, Rubtsova K, Scott-Browne J, Crawford F, Macdonald WA, et al. Crossreactive T cells spotlight the germline rules for alphabeta T cell-receptor interactions with MHC molecules. *Immunity.* (2008) 28:324–34. doi: 10.1016/j.immuni.2008.01.008
48. Feng D, Bond CJ, Ely LK, Maynard J, Garcia KC. Structural evidence for a germline-encoded T cell receptor-major histocompatibility complex interaction 'codon'. *Nat Immunol.* (2007) 8:975–83. doi: 10.1038/ni1502
49. Potter TA, Hansen TH, Habbersett R, Ozato K, Ahmed A. Flow microfluorometric analysis of H-2L expression. *J Immunol.* (1981) 127:580–4.
50. Matthews PC, Prendergast A, Leslie A, Crawford H, Payne R, Rousseau C, et al. Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. *J Virol.* (2008) 82:8548–59. doi: 10.1128/JVI.00580-08

51. Morishima Y, Kashiwase K, Matsuo K, Azuma F, Morishima S, Onizuka M, et al. Biological significance of HLA locus matching in unrelated donor bone marrow transplantation. *Blood.* (2015) 125:1189. doi: 10.1182/blood-2014-10-604785
52. Silberman D, Krovi SH, Tuttle KD, Crooks J, Reisdorph R, White J, et al. Class II major histocompatibility complex mutant mice to study the germ-line bias of T-cell antigen receptors. *Proc Natl Acad Sci USA.* (2016) 113:E5608–17. doi: 10.1073/pnas.1609717113
53. Gao K, Chen L, Zhang Y, Zhao Y, Wan Z, Wu J, et al. Germline-encoded TCR-MHC contacts promote TCR V gene bias in umbilical cord blood T cell repertoire. *bioRxiv.* (2019) 621821. doi: 10.1101/621821

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