# Human thymic T cell repertoire is imprinted with strong convergence to shared sequences

Heikkilä, Nelli

2020-11

Heikkilä , N , Vanhanen , R , Yohannes , D A , Kleino , I , Mattila , I P , Saramäki , J & Arstila , T P 2020 , ' Human thymic T cell repertoire is imprinted with strong convergence to shared sequences ' , Molecular Immunology , vol. 127 , pp. 112-123 . https://doi.org/10.1016/j.molimm.2020.09.003

http://hdl.handle.net/10138/340788
https://doi.org/10.1016/j.molimm.2020.09.003

# Human thymic T cell repertoire is imprinted with strong convergence to shared sequences

Heikkilä Nelli[a], Vanhanen Reetta[b], Yohannes Dawit A.[c], Kleino Iivari[d], Mattila Ilkka P.[e], Saramäki Jari[f], Arstila T. Petteri[g]

a) Research Programs Unit, Translational Immunology and Medicum, Department of Bacteriology and Immunology, University of Helsinki. Haartmaninkatu 3, 00290 Helsinki, Finland. nelli.heikkila@helsinki.fi
b) Research Programs Unit, Translational Immunology and Medicum, Department of Bacteriology and Immunology, University of Helsinki. Haartmaninkatu 3, 00290 Helsinki, Finland. reetta.vanhanen@helsinki.fi
c) Research Programs Unit, Translational Immunology and Medicum, Department of Medical and Clinical Genetics, University of Helsinki. Haartmaninkatu 8, 00290 Helsinki, Finland. dawit.yohannes@helsinki.fi
d) Research Programs Unit, Translational Immunology, University of Helsinki. Haartmaninkatu 3, 00290 Helsinki, Finland. iivari.kleino@helsinki.fi
e) Department of Pediatric Cardiac and Transplantation Surgery, Hospital for Children and Adolescents, Helsinki University Central Hospital. Stenbäckinkatu 9, 00290 Helsinki, Finland. ilkka.mattila@hus.fi
f) Department of Computer Science, Aalto University. Konemiehentie 2, 02150 Espoo, Finland. jari.saramaki@aalto.fi
g) Research Programs Unit, Translational Immunology and Medicum, Department of Bacteriology and Immunology, University of Helsinki. Haartmaninkatu 3, 00290 Helsinki, Finland. petteri.arstila@helsinki.fi

Corresponding author:
Nelli Heikkilä
Research Programs Unit, Translational Immunology, PB21, 00014 University of Helsinki, Finland
e-mail: nelli.heikkila@helsinki.fi
telephone: +358 29 41 26373
fax: +358 29 41 26382

Declarations of interest: none.

## Abstract

A highly diverse repertoire of T cell antigen receptors (TCR) is created in the thymus by recombination of gene segments and the insertion or deletion of nucleotides at the junctions. Using next-generation TCR sequencing we define here the features of recombination and selection in the human TCRα and TCRβ locus, and show that a strikingly high proportion of the repertoire is shared by unrelated individuals. The thymic TCRα nucleotide repertoire was more diverse than TCRβ, with $4.1 \times 10^6$ vs. $0.81 \times 10^6$ unique clonotypes, and contained nonproductive clonotypes at a higher frequency (69.2% vs. 21.2%). The convergence of distinct nucleotide clonotypes to the same amino acid sequences was higher in TCRα than in TCRβ repertoire (1.45 vs. 1.06 nucleotide sequences per amino acid sequence in thymus). The gene segment usage was biased, and generally all individuals favored the same genes in both TCRα and TCRβ loci. Despite the high diversity, a large fraction of the repertoire was found in more than one donor. The shared fraction was bigger in TCRα than TCRβ repertoire, and more common in in-frame sequences than in nonproductive sequences. Thus, both biases in rearrangement and thymic selection are likely to contribute to the generation of shared repertoire in humans.

**Keywords:** T cell antigen receptor, TCR repertoire, TCR recombination, thymus, next-generation sequencing

Abbreviations: T cell antigen receptor (TCR), V (variable), D (diversity), J (joining), CDR3 (complementarity-determining region 3), Pgen (generative probability)

## 1. Introduction

T cell antigen receptor (TCR) is a heterodimeric surface protein, consisting in most cells of α and β chains, while a small minority of cells use γ and δ chains. Both chains are encoded by genes assembled from incomplete segments via somatic recombination during development in the thymus. The TCRβ locus contains 47 variable (V), 2 diversity (D) and 13 joining (J) gene segments whereas the TCRα locus contains 42 V and 61 J segments but lacks the D segment. Further diversity is achieved at the gene segment junctions where a number of nucleotides may be removed and palindromic P-nucleotides and non-templated N-nucleotides inserted. Thus, most of the variability in the TCR concentrates in the junctional regions, called complementary determining region 3 (CDR3), which also form the main site of antigen recognition (Davis and Bjorkman, 1988).

The recombination process is capable of creating a high level of diversity. Direct sequencing of TCRβ repertoire has measured a lower limit of 1-3x10$^6$ clonotypes, whereas a mathematical estimator suggested a total repertoire of about 100x10$^6$ unique clonotypes (Qi et al., 2014). We have recently measured the lower limit of thymic TCR diversity in pediatric samples to be 10.3x10$^6$ for TCRβ and 3.7x10$^6$ for TCRα clonotypes, and statistical modelling suggested the total repertoire to consist of 40-70x10$^6$ and 60-100x10$^6$ clonotypes for TCRβ and TCRα respectively (Vanhanen et al., 2016). The pairing of TCRα to TCRβ has been studied but little. A sequencing of a limited TCR subset showed that on the average each TCRβ chain can bind to at least 24 different TCRα chains (Arstila et al., 1999), while a recent large scale single-cell analysis suggested that the pairing is more limited than would be compatible with a fully stochastic process (Grigaityte et al., 2017). The full TCRαβ repertoire thus consists of at least tens of millions of different receptors.

To date the human thymic TCR repertoire has been studied very little. The TCRβ locus is rearranged first and is subject to relatively stringent allelic exclusion. However, TCRβ locus may also be rearranged in cells destined to the γδ T cell lineage, which may account for a part of the nonfunctional TCRβ repertoire. Since most recombination events will result in an out-of-frame sequence, the functionality of the rearranged TCRβ chain is ensured by pairing with a surrogate TCRα chain, the preTα. The cells capable of signaling through pre-TCR then proliferate before recombination begins in the TCRα locus (von Boehmer et al., 1998). Unlike the TCRβ locus, in TCRα recombination both alleles are rearranged simultaneously, until a functional TCRαβ is expressed, stopping the recombination. Thus, in a large proportion of cells both TCRα loci are rearranged, although only one is likely to produce a functional protein chain (Casanova et al., 1991). The newly generated TCRαβ+ cells are then subjected to positive and negative selections, which remove cells incapable of interacting with HLA molecules or displaying too strong affinity to self-antigens (Klein et al., 2014). Overall, only an estimated 3-5% of the developing thymocytes survive the selection process to form the mature peripheral repertoire (Egerton et al., 1990; Yates, 2014).

In the present study, we characterize the composition of the thymic TCRβ and TCRα repertoire, identifying differences in the two chains related to their biology. Our data also show a strikingly strong convergence to shared repertoire in unrelated individuals.

## 2. Materials and Methods

The study was approved by the Pediatric Ethical Committee of the Helsinki University Hospital and parents gave a written informed consent. Thymus samples were obtained from eight immunologically healthy children undergoing a corrective operation for congenital cardiac defects (donors A-D and donors 1-4). Additionally, a peripheral blood sample was drawn from donors 1-4 during the operation. The donors were 7–244 days old and 2/8 were female (Table 1). Two of the subjects (donors A and B) were monozygotic twins. The impact of genetics on the repertoire has been analyzed in detail elsewhere (Heikkila et al., 2020). All thymus samples appeared macroscopically normal. Thymocyte populations from donors B-D were analyzed with flow cytometry for expression of CD4, CD8, TCRαβ, TCRγδ, CD3 and CD69.

From each subject, an aliquot of 10–30 million thymocytes and from donors 1-4 an aliquot of 0.5 mL peripheral blood was used for sequencing both TCRAD and TCRB repertoire. Thymocytes were extracted mechanically from the tissue. To remove red blood cells blood samples were treated with Gibco™ACK Lysing Buffer (Thermo Fisher Scientific, Massachusetts, USA), according to the manufacturer's orders. Genomic DNA was extracted from pelleted cell samples with QIAsymphony (Qiagen, Germany) according to the manufacturer's instructions. The TCRα and TCRβ CDR3 regions were amplified and sequenced from a standardized quantity of quality-controlled DNA using ImmunoSEQ assay (Adaptive Biotechnologies, Seattle, USA). In summary, the sequencing assay consists of a multiplex PCR system to amplify the rearranged CDR3 regions from the DNA samples at a length that is sufficient to subsequently identify the VDJ and VJ regions spanning each unique CDR3α and CDR3β regions, respectively. Amplicon sequencing was performed with Illumina platform. TCRα and TCRβ gene segment definitions were obtained from IMGT database (www.imgt.org). Primer bias was corrected as previously described (Vanhanen et al., 2016) and the resulting data filtered and clustered using both the relative frequency ratio between similar clones and a modified nearest-neighbor algorithm to remove both PCR and sequencing errors. All sequences are available at immuneACCESS database provided by Adaptive Biotechnologies (clients.adaptivebiotech.com/immuneaccess).

The TCR sequence analysis was performed using the immunoSEQ ANALYZER 3.0 (Adaptive Biotechnologies, Seattle, USA), VDJTools software (Shugay et al., 2015) and in-house scripts for computing languages R (www.r-project.org) and python 2.7 (www.python.org). The in-house scripts generated for this study are published in Supplements 1&2. The similarity of two sets of unique or total sequences was assessed calculating the Jaccard index, which is defined as the size of the intersection of two data sets (A and B) divided by the size of their union: $J(A,B) = \frac{|A \cup B|}{|A \cap B|}$. The abundance based Jaccard index was defined as $J_{abund} = UV/(U+V-UV)$, where U is the total relative abundance of shared sequences in sample A and V the total relative abundance of shared sequences in sample B (Chao et al., 2006). The CDR3 nucleotide sequences were extracted separately for in-frame and nonproductive sequences and subsequently the generative probabilities were calculated using the OLGA software (Sethna et al., 2019).

# 3. Results

## 3.1. TCRα and TCRβ repertoires differ in diversity and productivity

Thymus samples were collected from eight pediatric patients (donors A-D and donors 1-4), two of whom were monozygotic twins (donors A and B; Table 1). Flow cytometric analysis was performed for donors B-D and showed a normal distribution of CD4 and CD8 double-negative (DN), double-positive (DP), and single-positive (SP) thymocytes as well as normal pattern of TCRαβ and TCRγδ expression (Figure 1A). Postselection thymocytes were defined as DPCD3highCD69+, CD4SP or CD8SP (Swat et al., 1993; Yamashita et al., 1993). On the average, 23.1±3.7% of total thymocytes represented postselection and 76.9±3.7% preselection population (Figure 1B).

Sequencing of thymic TCRs yielded 1.2x10$^{5}$-1.6x10$^{6}$ (mean 810 000) unique TCRβ clonotypes of which 78.8±2.7 % were in-frame, 19.3±2.4 % were out-of-frame and 2.1±0.5 % contained a premature stop-codon (Fig. 1C). Consistent with our previous estimation on thymic TCR diversity, the TCRα diversity was higher than TCRβ diversity, with 1.3-7.6x10$^{6}$ (mean 4.1x10$^{6}$) unique clonotypes per sample (Vanhanen et al., 2016). However, the productivity in TCRα was much lower, as only 30.8±0.8 % of the unique clonotypes were in-frame. Of the unique TCRα clonotypes 66.0±0.6 % were out-of-frame and 7.0±4.5 % contained a premature stop-codon (Fig. 1C). As the sequencing assay is based on genomic DNA, it also provides a quantitative estimate of the number of total genomes with rearranged TCR segments in the sample.

A small blood sample from donors 1-4 was sequenced simultaneously with the thymus samples, producing an average of 84 000 unique TCRβ clonotypes and 150 000 unique TCRα clonotypes. In the TCRβ repertoire, the fractions of in-frame and nonproductive clonotypes remained essentially similar to that in the thymus (Figure 1D). In the TCRα repertoire, the fraction of in-frame clonotypes was higher in the blood samples than in the thymus (38.5±1.4% vs. 30.8±0.8%; Figure 1D).

To estimate the convergence of distinct nucleotide clonotypes to identical amino acid chains we calculated the nucleotide-to-amino acid-ratio for each sample. The majority of amino acid chains in the TCRβ repertoire were encoded by a single nucleotide clonotype, the nucleotide-to-amino acid-ratio being for unique in-frame clonotypes 1.06±0.03 in the thymus and 1.05±0.02 in the periphery. In the TCRα repertoire the number of unique nucleotide clonotypes converging to the same amino acid chain was higher than in the TCRβ repertoire, particularly in the thymus (ratio 1.45±0.13) but also to some degree in the periphery (ratio 1.18±0.01).

### 3.2. The V and J segment usage is biased before thymic selections

Previous studies of peripheral repertoire have shown a biased usage of V and J genes in healthy subjects. Similarly, in thymus the use of V gene elements was uneven, and the same segments were favored in each individual both in thymus and in blood (e.g. TRBV5-1, TRBV27-01 and TRAV21-1, TRAV29-1; Supplement 3). Similar findings were also obtained for J gene usage (Supplement 4). The biased V and J gene usage pattern was largely observed both in the in-frame and nonproductive repertoire, indicating that it is due the recombination process rather than selection (Figure 2). Consistent with our previous study (Heikkila et al., 2020), the samples from the monozygotic twins A and B clustered together, indicating a genetic component in V and J gene usage. Interestingly, in the TCRα repertoire, the gene segment usage clustered thymic and peripheral blood samples mainly according to the sample type and not the identity of the donor (Figure 2A). In the TCRβ repertoire, in contrast, the gene segment usage clustered together blood and thymus samples taken from the same donor (Figure 2B).

Some of the gene segment bias might be caused by thymic generation of semi-invariant T cell subsets, such as natural killer T cells (NKTs) or mucosal-associated T cells (MAITs). Human NKTs prefer TRAV10/TRAJ18 combination and MAITs use invariable TRAV01-02/TRAJ33-01 combination. The β chain usage is less restricted, but with a preference of TRBV25 for NKTs and TRBV6 and TRBV20 for MAITs. In our data none of the semi-invariant α chains was dominant whereas some MAIT-associated TRBV6 genes were found at an elevated frequency. However, these TRBV segments are also ubiquitously used by conventional variable T cells (Tickotsky et al., 2017).

TCRδ gene segments are embedded within the TCRα locus and αβ and γδ lymphocytes may use both TCRα and TCRδ gene segments in an overlapping manner (Verschuren et al., 1998). Since the thymocytes we analyzed were not sorted, and the sequencing protocol included primers specific for the entire TCRAD locus, we obtained a mixture of TCRα and TCRδ sequences. In the thymus, the frequency of γδ TCR+ thymocytes, as measured by flow cytometry, was 0.80±0.20%. However, the frequency of unique clonotypes using a combination of TRDV and TRDJ was 1.1±0.18%. In the peripheral blood the frequency of TRDV-TRDJ combinations was slightly higher (1.7±0.96 % of the unique clonotypes). We also identified relatively frequent combinations of TRDV to TRAJ (2.4±0.20 % of the unique clonotypes), whereas sequences using a combination of TRAV and TRDJ were rare both in thymus and in periphery (Table 2).

### 3.3. CDR3 region length reflects recombination and selection events

The TCRβ chain comprises V, D, and J segments, whereas the TCRα chain lacks D segments and thus contains only one junctional site. This difference was reflected in the higher number of non-templated nucleotide insertions in the TCRβ than in the TCRα sequences with an average of 9.3 vs. 3.9 nucleotides in thymic and 7.0 vs. 3.7 in peripheral in-frame repertoires (Figure 3A). The nonproductive sequences cannot be subject to TCR-mediated selection, and thus represent the non-selected product of the recombination process. Consistent with the previously reported shortening of CDR3 during thymic selection (Matsutani et al., 2011; Niemi et al., 2015; Yassai and Gorski, 2000), the mean CDR3 length was shorter in the in-frame rearrangements (41.6 base pairs (bp) for TCRα and 45.7 bp for TCRβ) than in the nonproductive rearrangements (41.9 bp for TCRα and 46.3 bp for TCRβ) in the thymus (Figure 3B). In the peripheral in-frame repertoire, the CDR3 regions were still shorter (41.4 bp for TCRα and 43.4 bp for TCRβ).

### 3.4. Pgen distributions differ in TCRα and TCRβ repertoires

In the process of V(D)J gene segment recombination and insertion of random nucleotides between gene segments some sequences are generated more readily while the generation of others is more unlikely. We used OLGA software to calculate the generative probabilities (Pgen) in the TCRα and TCRβ repertoires (Sethna et al., 2019). For a large majority of nonproductive sequences we obtained Pgen values 0, probably because the OLGA calculations are based on amino acid rather than nucleotide sequences and CDR3 amino acid definition remains ambivalent for nonproductive sequences. For the thymic in-frame sequences the Pgen was higher for TCRα (average Pgen 1.57e-7) than for TCRβ (average Pgen 1.34e-9) repertoire, a finding likely due to the lower junctional complexity in TCRα chains. The same was observed in the peripheral repertoires (average Pgen for TCRα 1.56e-7 and for TCRβ 3.62e-9). In the TCRα repertoire, the thymic and peripheral Pgen averages and distributions were largely identical, while for the TCRβ the thymic repertoires had lower Pgen values than the peripheral repertoires (1.34e-9 vs. 3.62e-9; Figure 4).

### 3.5. Overlap of thymic clonotypes between two individuals

Despite the high diversity of the junctional CDR3 sequences, a considerable overlap of peripheral TCR repertoires between different individuals has been reported (Shugay et al., 2013). In our thymic samples, a substantial fraction of TCR sequences were shared between two individuals, and some of the TCRα and TCRβ clonotypes were shared even between multiple individuals (Figure 5A). This phenomenon was more marked in the TCRα than TCRβ repertoire. Indeed, in the samples 1-4, in which the sequencing depth was shallower, there were no TCRβ clonotypes shared by all four donors.

To estimate the fraction of thymic repertoire shared by two individuals, we used the Jaccard index (JI), calculated as the intersection of two samples divided by the union of the samples, with a maximum index of 1 for fully overlapping repertoires. In the nonproductive TCRβ clonotypes the JI was low (mean JI 6.3e-5), but increased clearly in the in-frame repertoire (mean JI 4.6e-4). When unique amino acid CDR3 regions were analyzed, the shared fraction was higher still (mean JI 0.013; Figure 5B). In the TCRα repertoire, the shared fraction was generally higher than in the TCRβ repertoire, and in the nonproductive clonotypes the mean JI was 0.029. A small but consistent increase to mean JI of 0.032 was found in the in-frame repertoire. In the unique amino acid CDR3 regions the shared fraction was again clearly higher (mean JI 0.10; Figure 5B). As previously reported (Heikkila et al. 2020), comparison of the twins A and B produced slightly higher JIs than the other pairs. In general, samples 1-4 were sequenced to a lesser depth than samples A-D, affecting the observed number of shared clonotypes, and the JI values were consequently smaller. However, the increasing trend in JI from nonproductive to in-frame and amino acid sequences was clear in all samples.

The shared sequences contained fewer non-templated insertions than the individual private repertoires. The average number of non-templated insertions in was 1.4 and 2.6 respectively for shared in-frame and nonproductive TCRβ clonotypes. In TCRα the shared in-frame clonotypes contained on the average 1.4 and nonproductive 1.6 insertions. Also, the Pgen was higher in the shared repertoire compared to the full repertoires. In the in-frame repertoire the average Pgen for unique shared in-frame TCRβ clonotypes was 4.71e-8 and in the full repertoire 1.34e-9. For in-frame TCRα clonotypes the difference in Pgen between shared (2.38e-7) and full (1.57e-7) repertoires was smaller than for TCRβ but still distinct.

Since our sequencing method uses genomic DNA instead of messenger-RNA as starting material, it has been optimized for quantitative analysis and provides us with a reasonable estimate of the clonal abundance (Robins et al., 2009; Vanhanen et al., 2016). Thus, the analysis of the shared fraction of total genomes reflects the actual size of repertoire common to different individuals. For total genomes, a similar increasing trend in JIs from nonproductive to in-frame and to amino acid repertoires was observed as seen for unique sequences. In total in-frame nucleotide genomes the mean JI for TCRα repertoire was 0.083 and for TCRβ repertoire 0.00063. In total amino acids the shared part of the repertoire was extremely large (mean JI 0.30 for TCRα and 0.026 for TCRβ; Figure 5C). In percentages, on the average, of the total TCRβ amino acid repertoire of any given individual 6.1% was also found in the repertoire of another individual (range 1.55-11.4%). In the TCRα repertoire the overlap in percentages was strikingly high (mean 46.7%, range 32.6-62.7%; Supplement 5).

### 3.6. Sharing of high abundance clones

To analyze the relationship between clone size and the likelihood of sharing, we calculated the Jaccard indexes for the most abundant 1%, 2%, 5%, 10%, 20% and 50% of clones. For this analysis samples 1-4 were excluded, because the relatively shallow sequencing produced very little overlap among the top 1-5% clonotypes. In TCRα repertoire we observed a clear correlation between the sharing and the clonotype abundance. JI values were clearly highest in the top 1-2% most abundant clonotypes and decreased gradually when less abundant clonotypes were included (Figure 6A). In contrast, there was no similar correlation in the TCRβ repertoire and the interindividual variation in JIs among the top 1% most abundant clones was very wide (Figure 6A). The number of non-templated nucleotide inserts also showed a correlation with the sharing among highly abundant clones. Non-templated inserts were rare among the most abundant shared clones. In the TCRα repertoire the average number of inserts in the shared repertoire increased steadily with the analysis of less abundant clonotypes (Figure 6B). In the TCRβ repertoire the number of inserts was typically zero among the top 2% most abundant shared repertoire and increased abruptly for the top 5-50% most abundant clonotypes (Figure 6B).

### 3.7. Sequence overlap in the peripheral samples

Despite the clearly smaller number of cells analyzed, clonotype sharing was also observed in the peripheral blood. Similarly to the thymus, sharing was higher in the TCRα than in the TCRβ repertoire and some clonotypes were shared between all four samples (Figure 7A). Also in the peripheral samples, sharing was lowest in the nonproductive nucleotide repertoire, increased in the in-frame nucleotide and even more so in the amino acid CDR3 repertoires (Figure 7B-C).

## 4. Discussion

Until recently, our understanding of the human thymus has been largely based on extrapolation from circulating repertoire and from murine studies. However, studies on organ donors combined with high-throughput techniques and next-generation sequencing have begun to provide information on the various types of cells in the human thymus (Park et al., 2020; Thome et al., 2016). A single-cell sequencing study coupled with TCRαβ profiling identified approximately 200 000 individual lymphoid cells among 24 fetal and mature thymi and showed a biased V(D)J usage originating from recombination and modified by selection (Park et al., 2020). We have previously estimated the total thymic TCR diversity to be 60-100x$10^6$ for TCRα and 40-70x$10^6$ for TCRβ repertoire and thus currently beyond the coverage of single-cell experiments (Vanhanen et al., 2016). Our current data from eight pediatric thymi comprises a total of 161 million TCRα reads and 55 million TCRβ reads, representing the most extensive characterization of the thymic TCR repertoire so far. Although our analysis was performed on unsorted cells and thus allows little conclusions on the developmental stage and functionality of the TCRs, the large scale provides an opportunity to compare specific features of TCRα and TCRβ repertoires and, particularly, to measure thymic repertoire overlap across individuals.

As previously reported for peripheral blood samples and recently for thymus as well (Park et al., 2020; Quiros Roldan et al., 1995; Zvyagin et al., 2014) the usage of V and J gene segments is clearly biased in the thymus. The same gene segments were dominant in every individual, in both the TCRα and TCRβ chains. Some of this bias has been ascribed to selection by HLA molecules, which interact with protein loops encoded by the germ-line parts of TCR V genes (Huseby et al., 2005; Rudolph et al., 2006; Wu et al., 2002). However, the same biased usage was also observed in the nonproductive repertoire, which cannot be subjected to selection by antigen-HLA complexes. This suggests that the bias is partly generated in the recombination itself. We have previously reported that genetic factors influence the gene segment usage in the thymus, a finding confirmed here with an increased number of samples. Our data also show that the use of TCRD elements in αβ T cells is common, with ca. 6% of thymic sequences containing TCRD gene segments, while the frequency of γδ TCR+ thymocytes was less than 1%. However, combining TCRAV to TCRDJ seems to be largely prevented.

Despite the structural and functional similarity of the two TCR chains, the generation of TCRα and TCRβ repertoire has several differences, which are also reflected in our data. First, the number of non-templated nucleotide insertions was much higher in the TCRβ locus, most likely explainable by the fact that, unlike TCRα chain, TCRβ chain undergoes two recombination events (D to J followed by V to DJ). This is also displayed by the slightly longer CDR3 region length and lower calculated Pgen in TCRβ than in TCRα repertoire. Second, the number of non-templated inserts and CDR3 length were lower and respectively Pgen was higher in the peripheral than in thymic samples in TCRβ repertoire while in TCRα these features remained relatively similar in thymus and periphery. Third, the fraction of in-frame rearrangements was higher in TCRβ than in TCRα locus (78.8% vs. 30.8% in the thymus). This reflects the difference in allelic exclusion in TCRβ and TCRα locus. In TCRβ locus the exclusion is strict, whereas both TCRα loci are rearranged simultaneously and a large fraction of cells will end up with a nonfunctional rearrangement in the other TCRα locus (Borgulya et al., 1992; Casanova et al., 1991). The frequency of nonproductive sequences in our samples is also increased by the presence of immature thymocytes not yet subjected to TCR-mediated selection, and the ongoing TCRα locus rearrangement in some of the cells. Furthermore, the repertoire overlap was much higher in TCRα than TCRβ repertoire, consistent with previous analyses (Khosravi-Maharlooei et al., 2019; Zvyagin et al., 2014), but here shown in a large-scale analysis of thymic repertoire. Although we obtained fewer TCRβ than TCRα sequences, it is clear that the higher sequence overlap in TCRα compared with TCRβ is mostly biological and not due to differences in sequencing depth, a finding also confirmed by others.

Indeed, the remarkably high degree of clonal sharing between individuals is the most interesting observation in our current data. Here, it must be noted that two of our donors were monozygotic twins, which introduces a bias to the analysis. However, these two samples only shared a marginally higher fraction of sequences than the unrelated samples (Heikkila et al., 2020). This is largely consistent with a recent analysis of the peripheral repertoire in three pairs of identical twins, which concluded that there was no difference between the twins and unrelated donors in the sharing of CDR3 sequences (Zvyagin et al., 2014).

Given the enormous diversity of possible TCRs, the expected likelihood to detect identical receptors in two individuals is practically nonexistent. Still, previous studies of inbred mouse lines have reported that roughly 30% of the peripheral TCRβ repertoire is shared (Bousso et al., 1998; Furmanski et al., 2008). More recent studies have used next-generation sequencing methods, analyzing much larger numbers of sequences. Sequencing of TRBV12-4/TRBJ1-2 expressing peripheral blood CD8+ T cells in four unrelated healthy donors yielded in average 29 000 unique clonotypes per individual and the overlap of unique amino acid CDR3 sequences was 3.8–9.8% (Venturi et al., 2011). Zvyagin et al. measured the overlap of both TCRβ and TCRα repertoires in three pairs of monozygotic twins reaching an overlap of 3–10% and 10–26.5% of unique amino acid CDR3 clonotypes in TCRβ and TCRα repertoires, respectively, without higher similarity between the twins than unrelated pairs (Zvyagin et al., 2014). It was also estimated that if the predicted peripheral TCRβ diversity of $5 \mathrm{x} 10^6$ unique sequences was entirely sequenced, the CDR3 overlap between two individuals would reach 44.1% in the amino acid and 3.6% in the nucleotide repertoire (Shugay et al., 2013).

In our thymus data, taking into account the clonal abundances of the clonotypes, the average fraction of sequences found in any other donor for the total TCRβ repertoire was 0.2% for in-frame nucleotide chains and 6.1% for amino acid CDR3 chains. In the TCRα repertoire, the sharing was much higher: an average of 15.7% total in-frame nucleotide and 46.7% total amino acid CDR3 chains were shared between any two donors. Furthermore, in the TCRα repertoire the overlap showed a strong correlation with clone abundance, whereas the same was not true of the TCRβ repertoire. The average number of non-templated inserts was also lower among the most abundant clonotypes both for TCRα and TCRβ repertoire. Together, these data suggest that some TCRα sequences are generated easily and preferred across different unrelated individuals.

Notably, the surprisingly high repertoire overlap in thymus was directly measured from samples of 10 million thymocytes, taken from an organ with an estimated 50 billion cells (Ganusov and De Boer, 2007; Rodewald, 2008). Recent analyses have shown that the degree of sharing is correlated with the size of the sample sequenced (Campregher et al., 2010; Putintseva et al., 2013; Shugay et al., 2013; Venturi et al., 2011). It is thus possible that exhaustive sequencing of the thymic repertoire would reveal an even higher proportion of shared sequences. Indeed, it may be that TCR repertoire is really individualized only by α-to-β pairing, although even this may be less stochastic than previously assumed (Grigaityte et al., 2017).

It is clear that some of this sharing reflects convergent recombination, i.e., the recombination machinery favoring certain gene segments and particular types of CDR3 sequences. Previous analysis of peripheral repertoire has shown that shared sequences have relatively few nucleotide additions and are generally closer to germline sequences (Pogorelyy et al., 2017; Quigley et al., 2010; Venturi et al., 2008a; Venturi et al., 2006). This is also seen in our thymus samples, where the shared sequences had on average fewer nucleotide insertions and higher calculated Pgen than the repertoire in general. This implies that some junctional sequences are easier to generate and therefore appear repeatedly, and their high frequency may therefore not require peripheral expansion (Venturi et al., 2008b).

However, our quantitation showed a strong enrichment of the shared repertoire the further the sequences receded from the recombination process. In every donor pair the shared fraction was higher in the in-frame than in the nonproductive repertoire and higher still in amino acid sequences and total number of genomes. This was particularly striking in the TCRβ repertoire, in which the average JI increased from $6.3 \times 10^{-5}$ in the nonproductive repertoire to 0.026 in total amino acid genomes, or by a factor of ~400. In the TCRα chain the increase was by a factor of ~10, from 0.026 to 0.30. Since the nonproductive nucleotide sequences are not subject to any form of TCR-mediated selection, this enrichment indicates that a substantial fraction of the clonal sharing is due to antigen-driven selection in the thymus.

In the periphery, although shared clones specific to defined antigens have been described, the antigen-dependent selection seems in general to lead to divergence in the repertoire. Analysis of naive and memory CD8+ T cells found fewer shared clones in the latter, antigen-experienced population, while a comparison of preterm neonates with adults showed that the shared fraction of TCRβ (CDR3 amino acid chains) decreased from 8% to 1% (Carey et al., 2017). Similarly, donors in younger age groups shared a larger fraction of TCRβ repertoire than older individuals and while TCRβ repertoires in young are similarly high in diversity, with age clonal expansions accumulate and the individual repertoires develop to divergent directions (Britanova et al., 2014; Britanova et al., 2016). In our data the shared fraction of CDR3 amino acid sequences in the peripheral blood was 5.3% in the TCRβ and 17.1% in the TCRα repertoire, in donors ranging from 7 days to 5 months of age.

A further point relates to the transitory nature of thymic function. Since thymus is a primary lymphoid organ constantly producing new T cells, any given clone will spend only a limited time in the thymus before either failing selection and dying or maturing and emigrating to periphery. The repertoire might thus also be expected to be transitory, with a different snapshot of the repertoire obtained at different points in time. In contrast, the high degree of interindividual clonal sharing suggests by extension that at different time points a given thymus is producing similar clones. Indirectly, our results imply that although the thymic T cell population and TCR repertoire is transitory, the clonal composition of human thymus is surprisingly stable.

In conclusion, our study provides the first detailed characterization of the human thymic TCRα and TCRβ repertoire, showing similarities and differences in the features of these two TCR chains. We also show an unexpectedly high overlap of thymic TCR repertoire between unrelated donors, especially in the TCRα chain. Moreover, our data indicate that this convergence is substantially driven by thymic selection. Finally, it must be noted that the specificity of any TCR is determined by α-to-β pairing, which our data do not address. As shown by Grigaityte et al., novel technology is finally allowing this part of the repertoire to be analyzed, as well (Grigaityte et al., 2017).

**Acknowledgements:** We thank Tamás Bazsinka for technical assistance. This study has been funded by University of Helsinki Research Foundation, Emil Aaltonen Foundation, The Finnish Medical Foundation, Biomedicum Helsinki Foundation, and MD PhD program of the University of Helsinki.

**Funding:** The funding sources were not involved in the study design, conceptualization or article preparation.

**Declaration of interest:** The authors declare no commercial or financial conflict of interest.

**References:**

**Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., Kourilsky, P.**, 1999. A direct estimate of the human alphabeta T cell receptor diversity. Science 286, 958-961

**Borgulya, P., Kishi, H., Uematsu, Y., von Boehmer, H.**, 1992. Exclusion and Inclusion of Alpha-T-Cell and Beta-T-Cell Receptor Alleles. Cell 69, 529-537. 10.1016/0092-8674(92)90453-J

**Bousso, P., Casrouge, A., Altman, J.D., Haury, M., Kanellopoulos, J., Abastado, J.P., Kourilsky, P.**, 1998. Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. Immunity 9, 169-178

**Britanova, O.V., Putintseva, E.V., Shugay, M., Merzlyak, E.M., Turchaninova, M.A., Staroverov, D.B., Bolotin, D.A., Lukyanov, S., Bogdanova, E.A., Mamedov, I.Z., Lebedev, Y.B., Chudakov, D.M.**, 2014. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J Immunol 192, 2689-2698. 10.4049/jimmunol.1302064

**Britanova, O.V., Shugay, M., Merzlyak, E.M., Staroverov, D.B., Putintseva, E.V., Turchaninova, M.A., Mamedov, I.Z., Pogorelyy, M.V., Bolotin, D.A., Izraelson, M., Davydov, A.N., Egorov, E.S., Kasatskaya, S.A., Rebrikov, D.V., Lukyanov, S., Chudakov, D.M.**, 2016. Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. J Immunol 196, 5005-5013. 10.4049/jimmunol.1600005

**Campregher, P.V., Srivastava, S.K., Deeg, H.J., Robins, H.S., Warren, E.H.**, 2010. Abnormalities of the alphabeta T-cell receptor repertoire in advanced myelodysplastic syndrome. Exp Hematol 38, 202-212. 10.1016/j.exphem.2009.12.004

**Carey, A.J., Hope, J.L., Mueller, Y.M., Fike, A.J., Kumova, O.K., van Zessen, D.B.H., Steegers, E.A.P., van der Burg, M., Katsikis, P.D.**, 2017. Public Clonotypes and Convergent Recombination Characterize the Naive CD8(+) T-Cell Receptor Repertoire of Extremely Preterm Neonates. Front Immunol 8, 1859. 10.3389/fimmu.2017.01859

**Casanova, J.L., Romero, P., Widmann, C., Kourilsky, P., Maryanski, J.L.**, 1991. T cell receptor genes in a series of class I major histocompatibility complex-restricted cytotoxic T lymphocyte clones specific for a Plasmodium berghei nonapeptide: implications for T cell allelic exclusion and antigen-specific repertoire. J Exp Med 174, 1371-1383

**Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.J.**, 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62, 361-371. 10.1111/j.1541-0420.2005.00489.x

**Davis, M.M., Bjorkman, P.J.**, 1988. T-cell antigen receptor genes and T-cell recognition. Nature 334, 395-402. 10.1038/334395a0

**Egerton, M., Scollay, R., Shortman, K.**, 1990. Kinetics of mature T-cell development in the thymus. Proc Natl Acad Sci U S A 87, 2579-2582

**Furmanski, A.L., Ferreira, C., Bartok, I., Dimakou, S., Rice, J., Stevenson, F.K., Millrain, M.M., Simpson, E., Dyson, J.**, 2008. Public T cell receptor beta-chains are not advantaged during positive selection. J Immunol 180, 1029-1039

**Ganusov, V.V., De Boer, R.J.**, 2007. Do most lymphocytes in humans really reside in the gut? Trends in Immunology 28, 514-518. 10.1016/j.it.2007.08.009

**Grigaityte, K., Carter, J.A., Goldfless, S.J., Jeffery, E.W., Hause, R.J., Jiang, Y., Koppstein, D., Briggs, A.W., Church, G.M., Vigneault, F., Atwal, G.S.**, 2017. Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire. bioRxiv.

**Heikkila, N., Vanhanen, R., Yohannes, D.A., Saavalainen, P., Meri, S., Jokiranta, T.S., Jarva, H., Mattila, I.P., Hamm, D., Sormunen, S., Saramaki, J., Arstila, T.P.**, 2020. Identifying the inheritable component of human thymic T cell repertoire generation in monozygous twins. Eur J Immunol 50, 748-751. 10.1002/eji.201948404

**Huseby, E.S., White, J., Crawford, F., Vass, T., Becker, D., Pinilla, C., Marrack, P., Kappler, J.W.**, 2005. How the T cell repertoire becomes peptide and MHC specific. Cell 122, 247-260. 10.1016/j.cell.2005.05.013

**Khosravi-Maharlooei, M., Obradovic, A., Misra, A., Motwani, K., Holzl, M., Seay, H.R., DeWolf, S., Nauman, G., Danzl, N., Li, H., Ho, S.H., Winchester, R., Shen, Y., Brusko, T.M., Sykes, M.**, 2019. Crossreactive public TCR sequences undergo positive selection in the human thymic repertoire. J Clin Invest 129, 2446-2462. 10.1172/JCI124358

**Klein, L., Kyewski, B., Allen, P.M., Hogquist, K.A.**, 2014. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat Rev Immunol 14, 377-391. 10.1038/nri3667

**Matsutani, T., Ogata, M., Fujii, Y., Kitaura, K., Nishimoto, N., Suzuki, R., Itoh, T.**, 2011. Shortening of complementarity determining region 3 of the T cell receptor alpha chain during thymocyte development. Mol Immunol 48, 623-629. 10.1016/j.molimm.2010.11.003

**Niemi, H.J., Laakso, S., Salminen, J.T., Arstila, T.P., Tuulasvaara, A.**, 2015. A normal T cell receptor beta CDR3 length distribution in patients with APECED. Cell Immunol 295, 99-104. 10.1016/j.cellimm.2015.03.005

**Park, J.E., Botting, R.A., Dominguez Conde, C., Popescu, D.M., Lavaert, M., Kunz, D.J., Goh, I., Stephenson, E., Ragazzini, R., Tuck, E., Wilbrey-Clark, A., Roberts, K., Kedlian, V.R., Ferdinand, J.R., He, X., Webb, S., Maunder, D., Vandamme, N., Mahbubani, K.T., Polanski, K., Mamanova, L., Bolt, L., Crossland, D., de Rita, F., Fuller, A., Filby, A., Reynolds, G., Dixon, D., Saeb-Parsy, K., Lisgo, S., Henderson, D., Vento-Tormo, R., Bayraktar, O.A., Barker, R.A., Meyer, K.B., Saeys, Y., Bonfanti, P., Behjati, S., Clatworthy, M.R., Taghon, T., Haniffa, M., Teichmann, S.A.**, 2020. A cell atlas of human thymic development defines T cell repertoire formation. Science 367. 10.1126/science.aay3224

**Pogorelyy, M.V., Elhanati, Y., Marcou, Q., Sycheva, A.L., Komech, E.A., Nazarov, V.I., Britanova, O.V., Chudakov, D.M., Mamedov, I.Z., Lebedev, Y.B., Mora, T., Walczak, A.M.**, 2017. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. PLoS Comput Biol 13. 10.1371/journal.pcbi.1005572

**Putintseva, E.V., Britanova, O.V., Staroverov, D.B., Merzlyak, E.M., Turchaninova, M.A., Shugay, M., Bolotin, D.A., Pogorelyy, M.V., Mamedov, I.Z., Bobrynina, V., Maschan, M., Lebedev, Y.B., Chudakov, D.M.**, 2013. Mother and child T cell receptor repertoires: deep profiling study. Front Immunol 4, 463. 10.3389/fimmu.2013.00463

**Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.Y., Olshen, R.A., Weyand, C.M., Boyd, S.D., Goronzy, J.J.**, 2014. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci U S A 111, 13139-13144. 10.1073/pnas.1409155111

**Quigley, M.F., Greenaway, H.Y., Venturi, V., Lindsay, R., Quinn, K.M., Seder, R.A., Douek, D.C., Davenport, M.P., Price, D.A.**, 2010. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. Proc Natl Acad Sci U S A 107, 19414-19419. 10.1073/pnas.1010586107

**Quiros Roldan, E., Sottini, A., Bettinardi, A., Albertini, A., Imberti, L., Primi, D.**, 1995. Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells. Immunogenetics 41, 91-100

**Robins, H.S., Campregher, P.V., Srivastava, S.K., Wacher, A., Turtle, C.J., Kahsai, O., Riddell, S.R., Warren, E.H., Carlson, C.S.**, 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114, 4099-4107. 10.1182/blood-2009-04-217604

**Rodewald, H.-R.**, 2008. Thymus Organogenesis. 10.1146/annurev.immunol.26.021607.090408

**Rudolph, M.G., Stanfield, R.L., Wilson, I.A.**, 2006. How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol 24, 419-466. 10.1146/annurev.immunol.23.021704.115658

**Sethna, Z., Elhanati, Y., Callan, C.G., Walczak, A.M., Mora, T.**, 2019. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics 35, 2974-2981. 10.1093/bioinformatics/btz035

**Shugay, M., Bagaev, D.V., Turchaninova, M.A., Bolotin, D.A., Britanova, O.V., Putintseva, E.V., Pogorelyy, M.V., Nazarov, V.I., Zvyagin, I.V., Kirgizova, V.I., Kirgizov, K.I., Skorobogatova, E.V., Chudakov, D.M.**, 2015. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLoS Comput Biol 11, e1004503. 10.1371/journal.pcbi.1004503

**Shugay, M., Bolotin, D.A., Putintseva, E.V., Pogorelyy, M.V., Mamedov, I.Z., Chudakov, D.M.**, 2013. Huge Overlap of Individual TCR Beta Repertoires. Front Immunol 4, 466. 10.3389/fimmu.2013.00466

**Swat, W., Dessing, M., von Boehmer, H., Kisielow, P.**, 1993. CD69 expression during selection and maturation of CD4+8+ thymocytes. Eur J Immunol 23, 739-746. 10.1002/eji.1830230326

**Thome, J.J., Bickham, K.L., Ohmura, Y., Kubota, M., Matsuoka, N., Gordon, C., Granot, T., Griesemer, A., Lerner, H., Kato, T., Farber, D.L.**, 2016. Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. Nat Med 22, 72-77. 10.1038/nm.4008

**Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., Friedman, N.**, 2017. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 33, 2924-2929. 10.1093/bioinformatics/btx286

**Vanhanen, R., Heikkila, N., Aggarwal, K., Hamm, D., Tarkkila, H., Patila, T., Jokiranta, T.S., Saramaki, J., Arstila, T.P.**, 2016. T cell receptor diversity in the human thymus. Mol Immunol 76, 116-122. 10.1016/j.molimm.2016.07.002

**Venturi, V., Chin, H.Y., Asher, T.E., Ladell, K., Scheinberg, P., Bornstein, E., van Bockel, D., Kelleher, A.D., Douek, D.C., Price, D.A., Davenport, M.P.**, 2008a. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. J Immunol 181, 7853-7862

**Venturi, V., Chin, H.Y., Price, D.A., Douek, D.C., Davenport, M.P.**, 2008b. The role of production frequency in the sharing of simian immunodeficiency virus-specific CD8+ TCRs between macaques. J Immunol 181, 2597-2609

**Venturi, V., Kedzierska, K., Price, D.A., Doherty, P.C., Douek, D.C., Turner, S.J., Davenport, M.P.**, 2006. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. Proc Natl Acad Sci U S A 103, 18691-18696. 10.1073/pnas.0608907103

**Venturi, V., Quigley, M.F., Greenaway, H.Y., Ng, P.C., Ende, Z.S., McIntosh, T., Asher, T.E., Almeida, J.R., Levy, S., Price, D.A., Davenport, M.P., Douek, D.C.**, 2011. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J Immunol 186, 4285-4294. 10.4049/jimmunol.1003898

**Verschuren, M.C., Wolvers-Tettero, I.L., Breit, T.M., van Dongen, J.J.**, 1998. T-cell receptor V delta-J alpha rearrangements in human thymocytes: the role of V delta-J alpha rearrangements in T-cell receptor-delta gene deletion. Immunology 93, 208-212

**von Boehmer, H., Aifantis, I., Azogui, O., Feinberg, J., Saint-Ruf, C., Zober, C., Garcia, C., Buer, J.**, 1998. Crucial function of the pre-T-cell receptor (TCR) in TCR beta selection, TCR beta allelic exclusion and alpha beta versus gamma delta lineage commitment. Immunol Rev 165, 111-119

**Wu, L.C., Tuot, D.S., Lyons, D.S., Garcia, K.C., Davis, M.M.**, 2002. Two-step binding mechanism for T-cell receptor recognition of peptide MHC. Nature 418, 552-556. 10.1038/nature00920

**Yamashita, I., Nagata, T., Tada, T., Nakayama, T.**, 1993. CD69 cell surface expression identifies developing thymocytes which audition for T cell antigen receptor-mediated positive selection. Int Immunol 5, 1139-1150. 10.1093/intimm/5.9.1139

**Yassai, M., Gorski, J.**, 2000. Thymocyte maturation: selection for in-frame TCR alpha-chain rearrangement is followed by selection for shorter TCR beta-chain complementarity-determining region 3. J Immunol 165, 3706-3712

**Yates, A.J.**, 2014. Theories and quantification of thymic selection. Front Immunol 5, 13. 10.3389/fimmu.2014.00013

**Zvyagin, I.V., Pogorelyy, M.V., Ivanova, M.E., Komech, E.A., Shugay, M., Bolotin, D.A., Shelenkov, A.A., Kurnosov, A.A., Staroverov, D.B., Chudakov, D.M., Lebedev, Y.B., Mamedov, I.Z.**, 2014. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. Proc Natl Acad Sci U S A 111, 5980-5985. 10.1073/pnas.1319389111

## Tables

**Table 1.** Description of the samples. The details of each sequenced sample and the numbers of obtained unique clonotypes and total reads per sample for TCRα and TCRβ repertoires.

| | | | TCRα | | TCRβ | |
|---|---|---|---|---|---|---|
| **Sample** | **Age (days)** | **Sex** | **Unique** | **Total** | **Unique** | **Total** |
| Thymus A | 243 | M | 6 907 422 | 39 865 283 | 1 254 760 | 8 431 833 |
| Thymus B | 244 | M | 7 578 104 | 45 335 572 | 1 540 161 | 11 558 445 |
| Thymus C | 225 | F | 5 347 824 | 30 309 225 | 1 568 528 | 23 581 729 |
| Thymus D | 126 | M | 6 743 495 | 36 762 724 | 1 462 150 | 11 159 872 |
| Thymus 1 | 7 | M | 2 089 557 | 3 179 774 | 223 725 | 237 063 |
| Thymus 2 | 52 | M | 1 262 845 | 1 747 487 | 173 368 | 182 356 |
| Thymus 3 | 107 | M | 1 289 728 | 2 158 043 | 138 544 | 142 903 |
| Thymus 4 | 156 | F | 1 419 013 | 1 848 851 | 122 195 | 128 228 |
| **Average** | | | **4 079 749** | **20 150 870** | **810 429** | **6 927 804** |
| Blood 1 | 7 | M | 138 159 | 154 682 | 77 868 | 82 418 |
| Blood 2 | 52 | M | 109 171 | 123 523 | 69 875 | 73 945 |
| Blood 3 | 107 | M | 180 100 | 245 126 | 104 236 | 134 110 |
| Blood 4 | 156 | F | 167 266 | 199 326 | 82 550 | 88 901 |
| **Average** | | | **148 674** | **180 664** | **83 632** | **94 844** |

**Table 2.** Mean frequency (%) of Vδ and Jδ segments in the TCRα repertoire

| | | Thymus | Peripheral blood |
|---|---|---|---|
| **Vδ-Jδ** | Unique | 1.09 | 1.71 |
| | Total | 2.06 | 3.83 |
| **Vδ-Jα** | Unique | 2.39 | 2.71 |
| | Total | 3.81 | 3.84 |
| **Vα-Jδ** | Unique | 0.36 | 0.42 |
| | Total | 0.41 | 0.51 |

## Figure captions

**Figure 1.** Analysis of the thymocyte subsets and repertoire productivity. The fraction of TCRαβ+ and TCRγδ+ in thymocytes and the distribution of CD4 and CD8 among TCRαβ+ thymocytes in a representative thymus sample (donor C) with the applied backgating (A). The distribution of CD4 and CD8 expression in thymocytes and the fraction of CD3$^{high}$CD69+ cells in CD4+CD8+ double positive thymocytes (donor C) with the applied backgating (B). The fraction of sequences in-frame, out-of-frame, or containing a premature stop codon among unique TCRα and TCRβ clonotypes for thymic (C) and peripheral TCR repertoires (D).

**Figure 2.** The V gene usage in in-frame and nonproductive repertoires. The heatmaps display the frequencies of different V gene segments and the attached dendrograms show the clustering of the samples in in-frame and nonproductive TCRα (A) and TCRβ repertoires (B).

**Figure 3.** The number of non-templated insertions and the CDR3 lengths. The graphs show the average and 95% confidence interval of the number of non-templated nucleotide insertions (A) and of CDR3 lengths (B) in thymic and peripheral blood TCRα and TCRβ repertoires for in-frame and nonproductive sequences.

**Figure 4.** The generation probability (Pgen) calculated with OLGA software. Thymic and peripheral Pgen distribution plotted against probability density in the in-frame TCRα and TCRβ repertoires for a representative thymus-blood pair (donor 1).

**Figure 5.** Sequence overlap between thymus samples. Venn diagrams show the overlap of unique in-frame clonotypes separately for thymus samples A-D and 1-4 (A). Individual Jaccard indexes

(JI) between each thymus sample for nonproductive, in-frame and amino acid repertoires among unique clonotypes (B) and total genomes (C). Monozygotic twins A and B are identified as open circles, filled circles represent the JI between unrelated individuals. The average JI and the 95% confidence interval are shown.

**Figure 6.** Sequence overlap among the high abundance clonotypes. Jaccard indexes (A) and non-templated insertions in the shared in-frame sequences (B) among the top 1%, 2%, 5%, 10%, 20% and 50% most abundant clonotypes and full repertoire in thymus samples A-D. The horizontal bars show the average and error bars indicate the 95% confidence interval.

**Figure 7.** Sequence overlap between peripheral blood samples. Venn diagrams show the overlap of unique in-frame clonotypes (A). Jaccard indexes (JI) for nonproductive, in-frame and amino acid repertoires among unique clonotypes (B) and total genomes (C). The average JI and the 95% confidence interval are displayed.

Figure 1

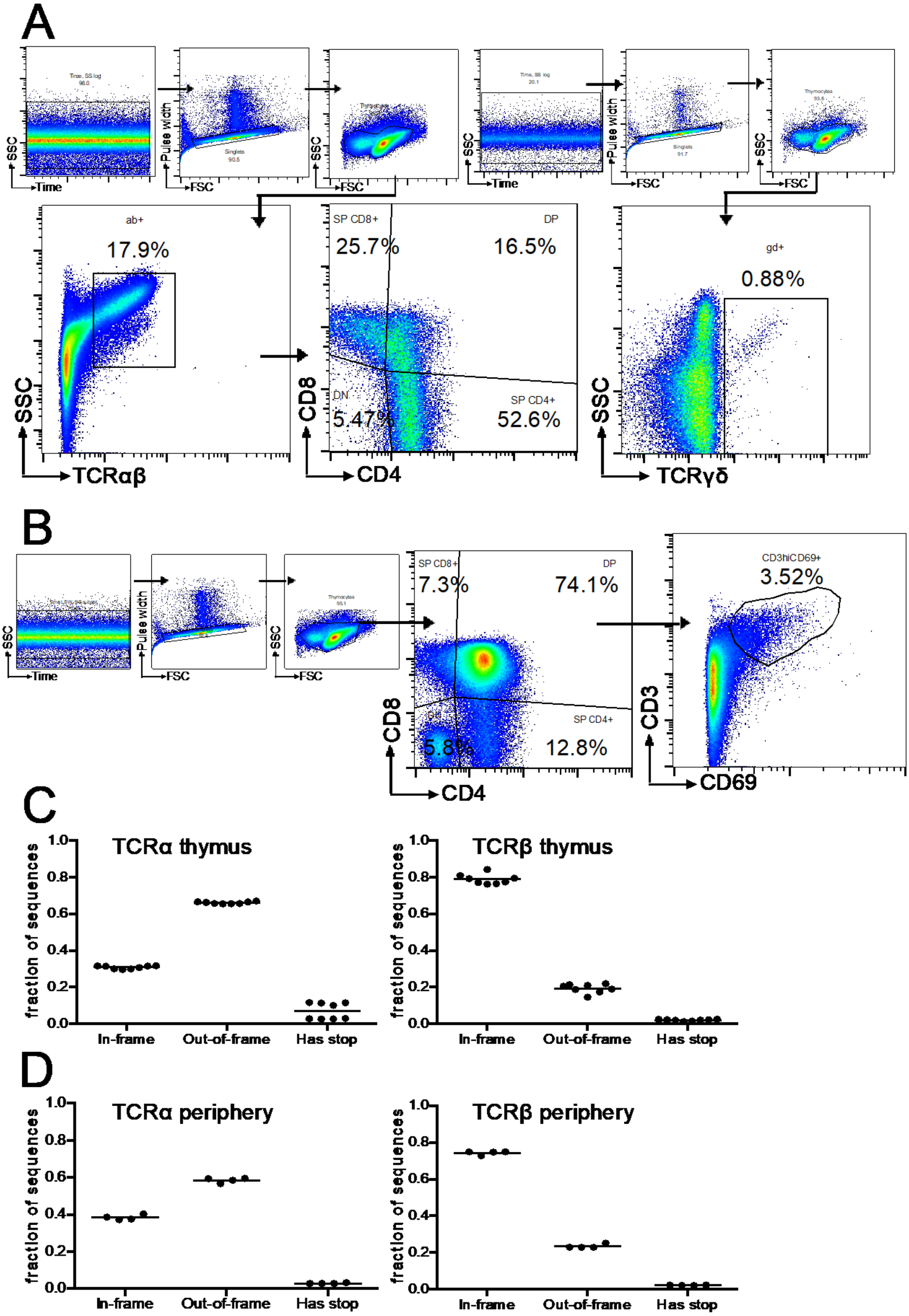
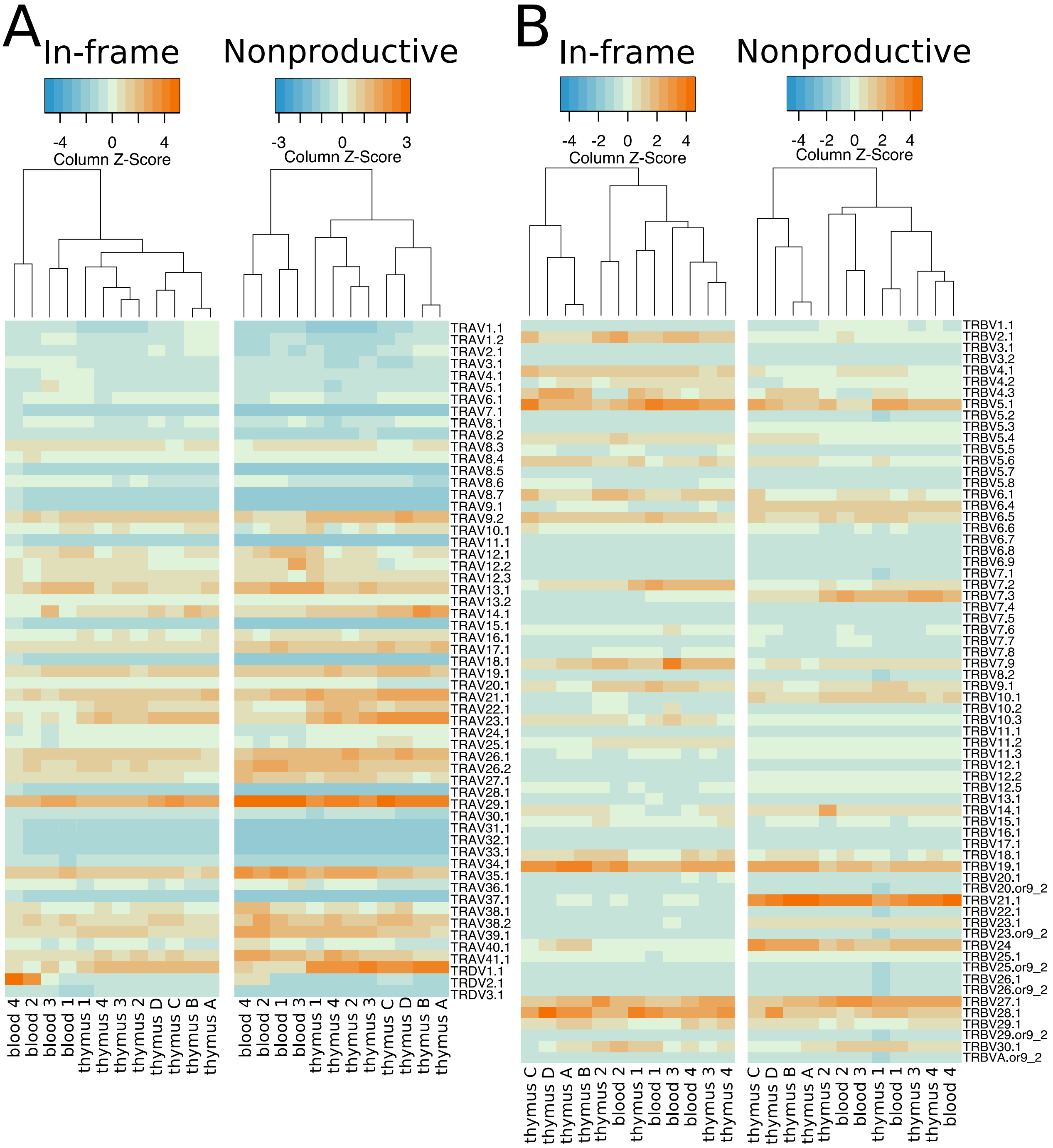
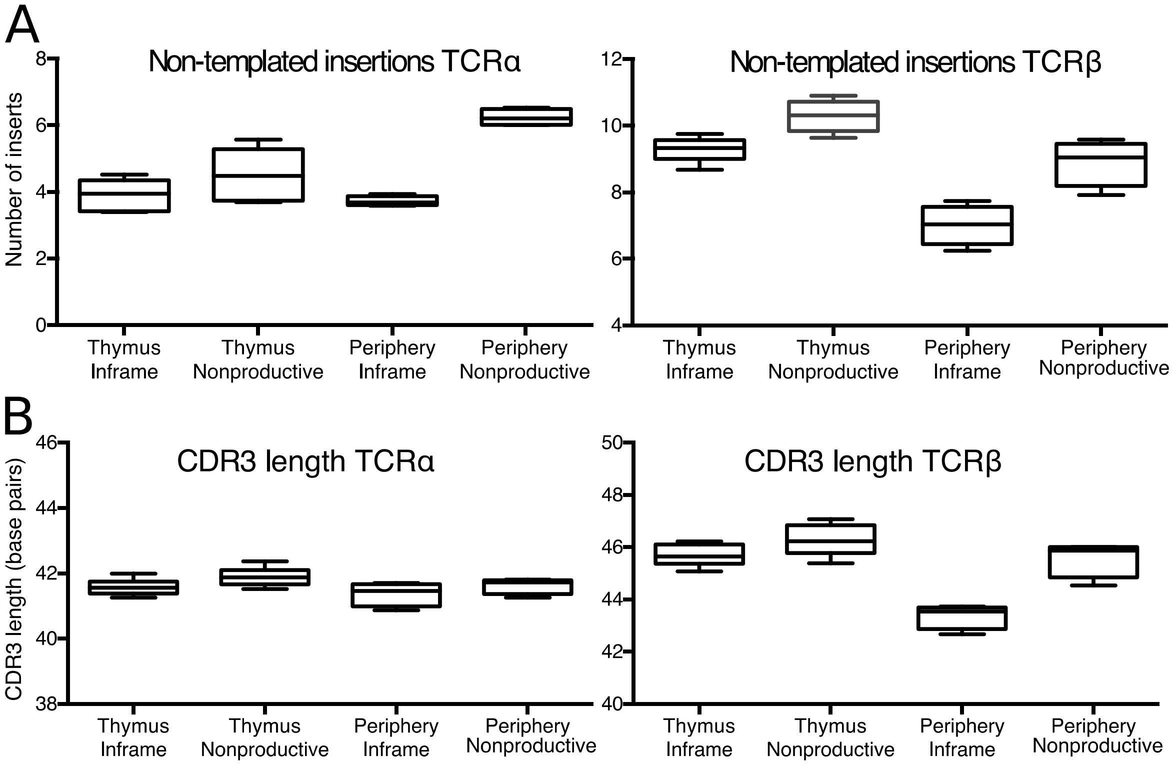

A
SSC
Time
Time, SS log
96.0
Pulse width
FSC
Singlets
90.5
SSC
FSC
Thymocytes
SSC
Time
Time, SS log
20.1
Pulse width
FSC
Singlets
91.7
SSC
FSC
Thymocytes
93.5
ab+
17.9%
SSC
TCRαβ
SP CD8+
25.7%
DP
16.5%
DN
5.47%
SP CD4+
52.6%
CD8
CD4
gd+
0.88%
SSC
TCRγδ
B
SSC
Time
Pulse width
FSC
SSC
FSC
Thymocytes
95.1
SP CD8+
7.3%
DP
74.1%
DN
5.8%
SP CD4+
12.8%
CD8
CD4
CD3hiCD69+
3.52%
CD3
CD69
C
TCRα thymus
fraction of sequences
1.0
0.8
0.6
0.4
0.2
0.0
In-frame
Out-of-frame
Has stop
TCRβ thymus
fraction of sequences
1.0
0.8
0.6
0.4
0.2
0.0
In-frame
Out-of-frame
Has stop
D
TCRα periphery
fraction of sequences
1.0
0.8
0.6
0.4
0.2
0.0
In-frame
Out-of-frame
Has stop
TCRβ periphery
fraction of sequences
1.0
0.8
0.6
0.4
0.2
0.0
In-frame
Out-of-frame
Has stop

Figure 2

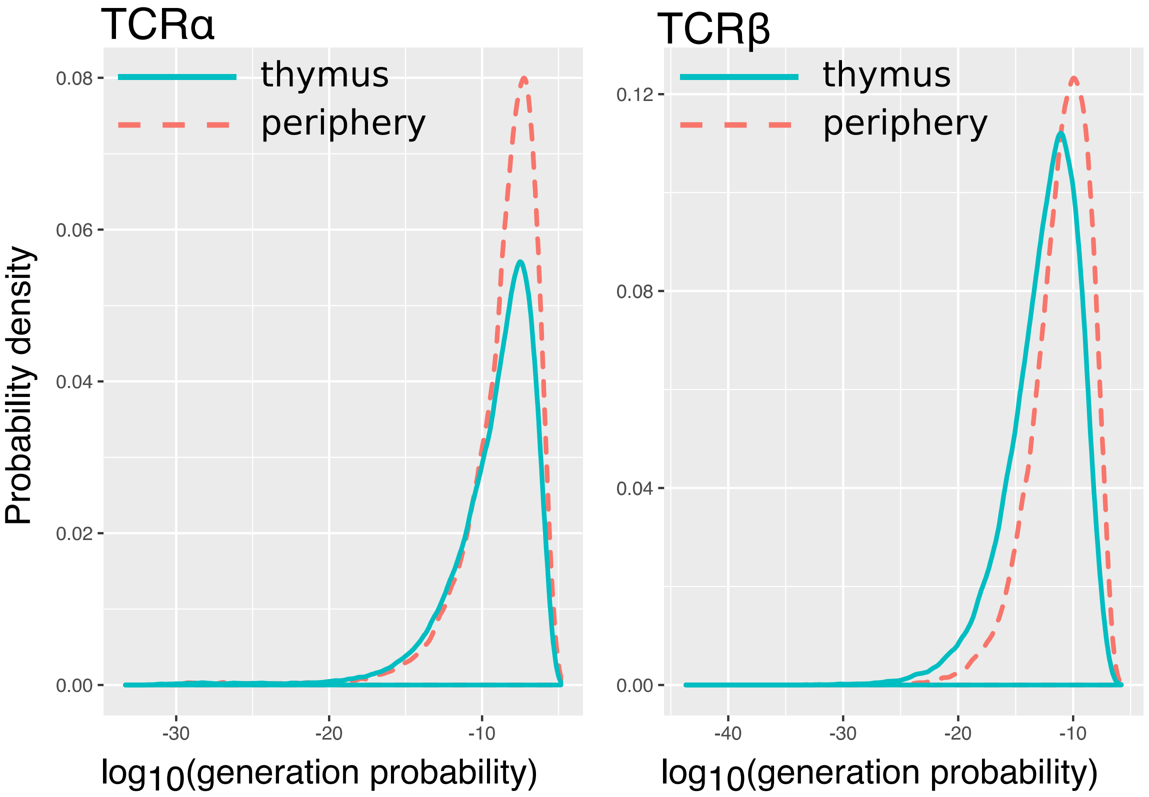

A
In-frame
-4 0 4
Column Z-Score
Nonproductive
-3 0 3
Column Z-Score
TRAV1.1
TRAV1.2
TRAV2.1
TRAV3.1
TRAV4.1
TRAV5.1
TRAV6.1
TRAV7.1
TRAV8.1
TRAV8.2
TRAV8.3
TRAV8.4
TRAV8.5
TRAV8.6
TRAV8.7
TRAV9.1
TRAV9.2
TRAV10.1
TRAV11.1
TRAV12.1
TRAV12.2
TRAV12.3
TRAV13.1
TRAV13.2
TRAV14.1
TRAV15.1
TRAV16.1
TRAV17.1
TRAV18.1
TRAV19.1
TRAV20.1
TRAV21.1
TRAV22.1
TRAV23.1
TRAV24.1
TRAV25.1
TRAV26.1
TRAV26.2
TRAV27.1
TRAV28.1
TRAV29.1
TRAV30.1
TRAV31.1
TRAV32.1
TRAV33.1
TRAV34.1
TRAV35.1
TRAV36.1
TRAV37.1
TRAV38.1
TRAV38.2
TRAV39.1
TRAV40.1
TRAV41.1
TRDV1.1
TRDV2.1
TRDV3.1
blood 4
blood 2
blood 3
blood 1
thymus 1
thymus 4
thymus 3
thymus 2
thymus D
thymus C
thymus B
thymus A
blood 4
blood 2
blood 1
blood 3
thymus 1
thymus 4
thymus 2
thymus 3
thymus C
thymus D
thymus B
thymus A
B
In-frame
-4 -2 0 2 4
Column Z-Score
Nonproductive
-4 -2 0 2 4
Column Z-Score
TRBV1.1
TRBV2.1
TRBV3.1
TRBV3.2
TRBV4.1
TRBV4.2
TRBV4.3
TRBV5.1
TRBV5.2
TRBV5.3
TRBV5.4
TRBV5.5
TRBV5.6
TRBV5.7
TRBV5.8
TRBV6.1
TRBV6.4
TRBV6.5
TRBV6.6
TRBV6.7
TRBV6.8
TRBV6.9
TRBV7.1
TRBV7.2
TRBV7.3
TRBV7.4
TRBV7.5
TRBV7.6
TRBV7.7
TRBV7.8
TRBV7.9
TRBV8.2
TRBV9.1
TRBV10.1
TRBV10.2
TRBV10.3
TRBV11.1
TRBV11.2
TRBV11.3
TRBV12.1
TRBV12.2
TRBV12.5
TRBV13.1
TRBV14.1
TRBV15.1
TRBV16.1
TRBV17.1
TRBV18.1
TRBV19.1
TRBV20.1
TRBV20.or9_2
TRBV21.1
TRBV22.1
TRBV23.1
TRBV23.or9_2
TRBV24
TRBV25.1
TRBV25.or9_2
TRBV26.1
TRBV26.or9_2
TRBV27.1
TRBV28.1
TRBV29.1
TRBV29.or9_2
TRBV30.1
TRBVA.or9_2
thymus C
thymus D
thymus A
thymus B
thymus 2
blood 2
thymus 1
blood 1
blood 3
blood 4
thymus 3
thymus 4
thymus C
thymus D
thymus B
thymus A
thymus 2
blood 2
blood 3
thymus 1
blood 1
thymus 3
thymus 4
blood 4

665 Figure 3



666

667 Figure 4



668

Figure 5

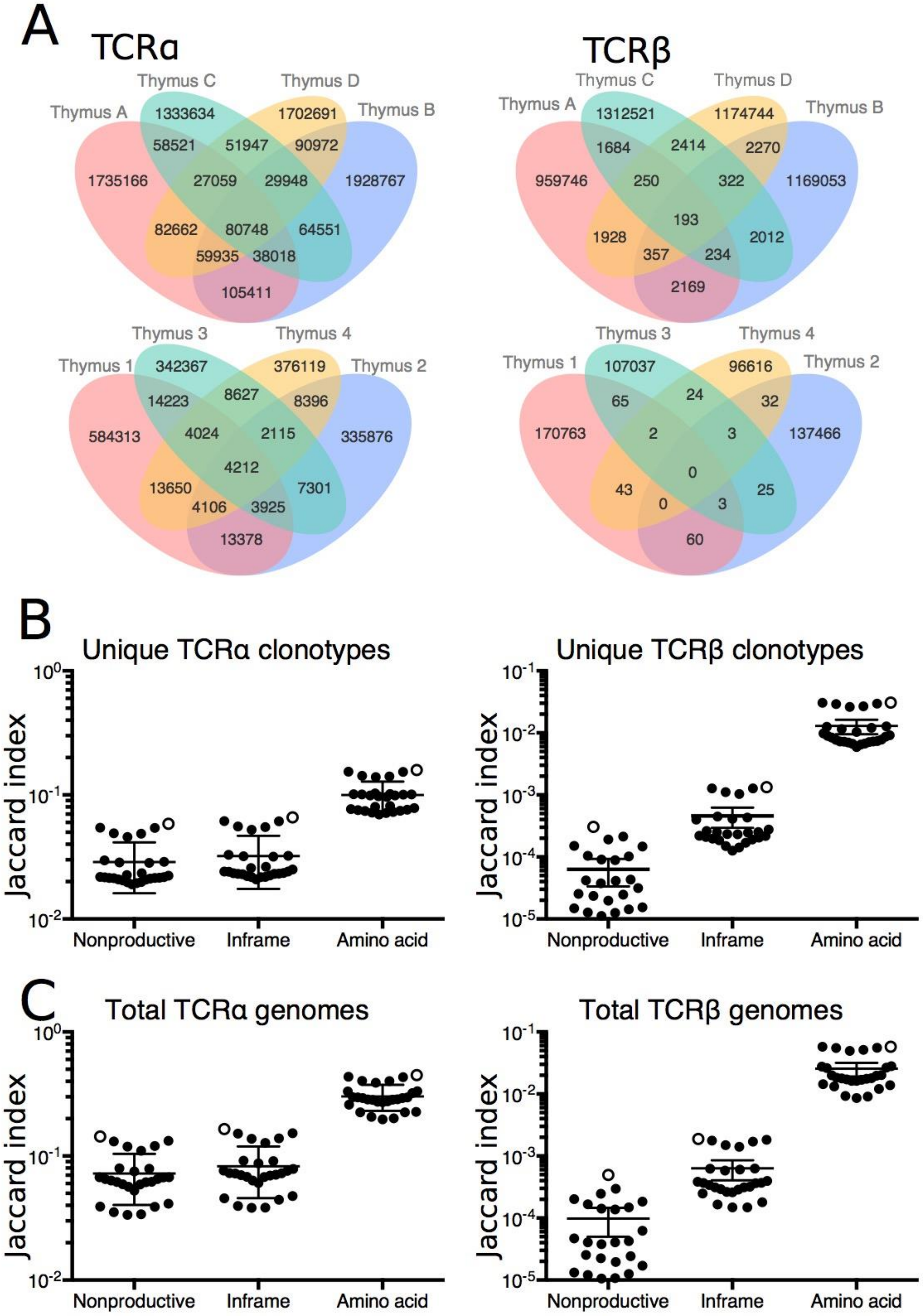

A
TCRα
Thymus A
Thymus C
Thymus D
Thymus B
1333634
1702691
58521
51947
90972
1735166
27059
29948
1928767
82662
80748
64551
59935
38018
105411
Thymus 1
Thymus 3
Thymus 4
Thymus 2
342367
376119
14223
8627
8396
584313
4024
2115
335876
4212
13650
7301
4106
3925
13378
TCRβ
Thymus A
Thymus C
Thymus D
Thymus B
1312521
1174744
1684
2414
2270
959746
250
322
1169053
193
1928
2012
357
234
2169
Thymus 1
Thymus 3
Thymus 4
Thymus 2
107037
96616
65
24
32
170763
2
3
137466
0
43
25
0
3
60
B
Unique TCRα clonotypes
Jaccard index
$10^{0}$
$10^{-1}$
$10^{-2}$
Nonproductive
Inframe
Amino acid
Unique TCRβ clonotypes
Jaccard index
$10^{-1}$
$10^{-2}$
$10^{-3}$
$10^{-4}$
$10^{-5}$
Nonproductive
Inframe
Amino acid
C
Total TCRα genomes
Jaccard index
$10^{0}$
$10^{-1}$
$10^{-2}$
Nonproductive
Inframe
Amino acid
Total TCRβ genomes
Jaccard index
$10^{-1}$
$10^{-2}$
$10^{-3}$
$10^{-4}$
$10^{-5}$
Nonproductive
Inframe
Amino acid

Figure 6

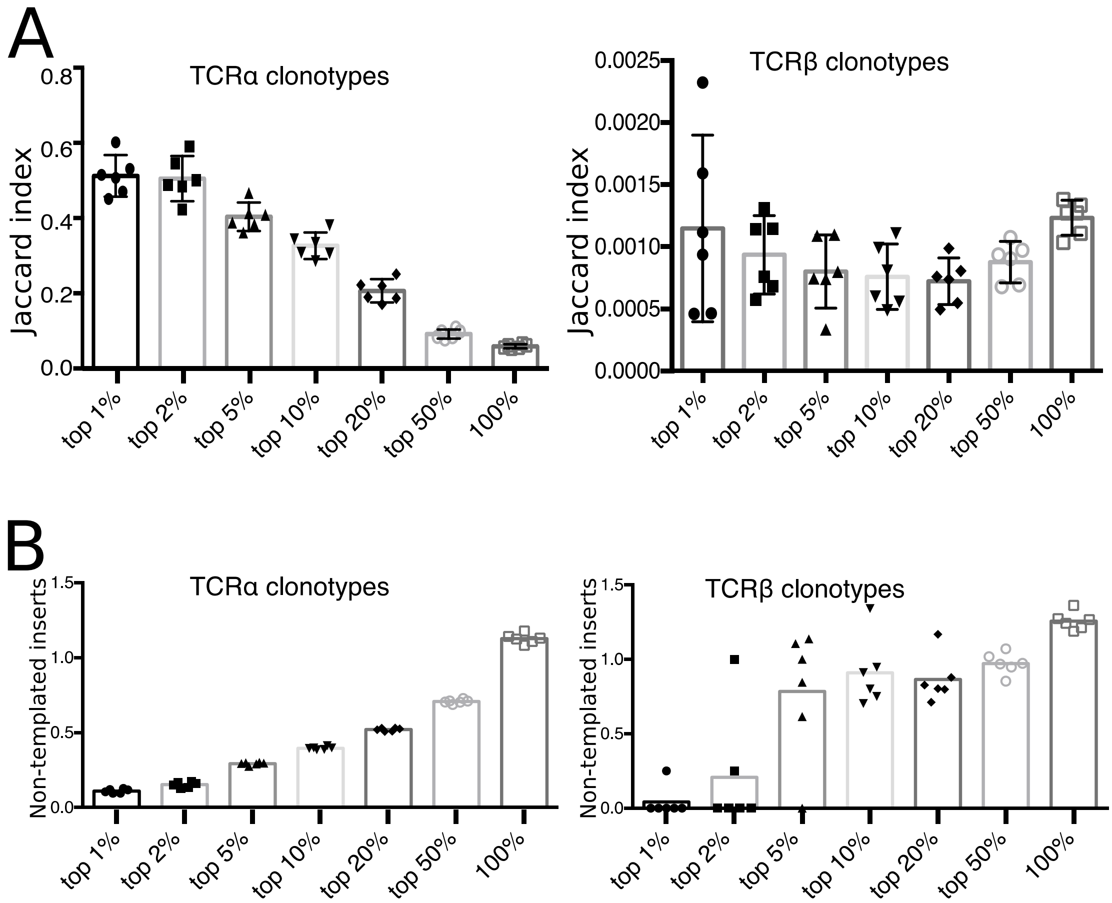
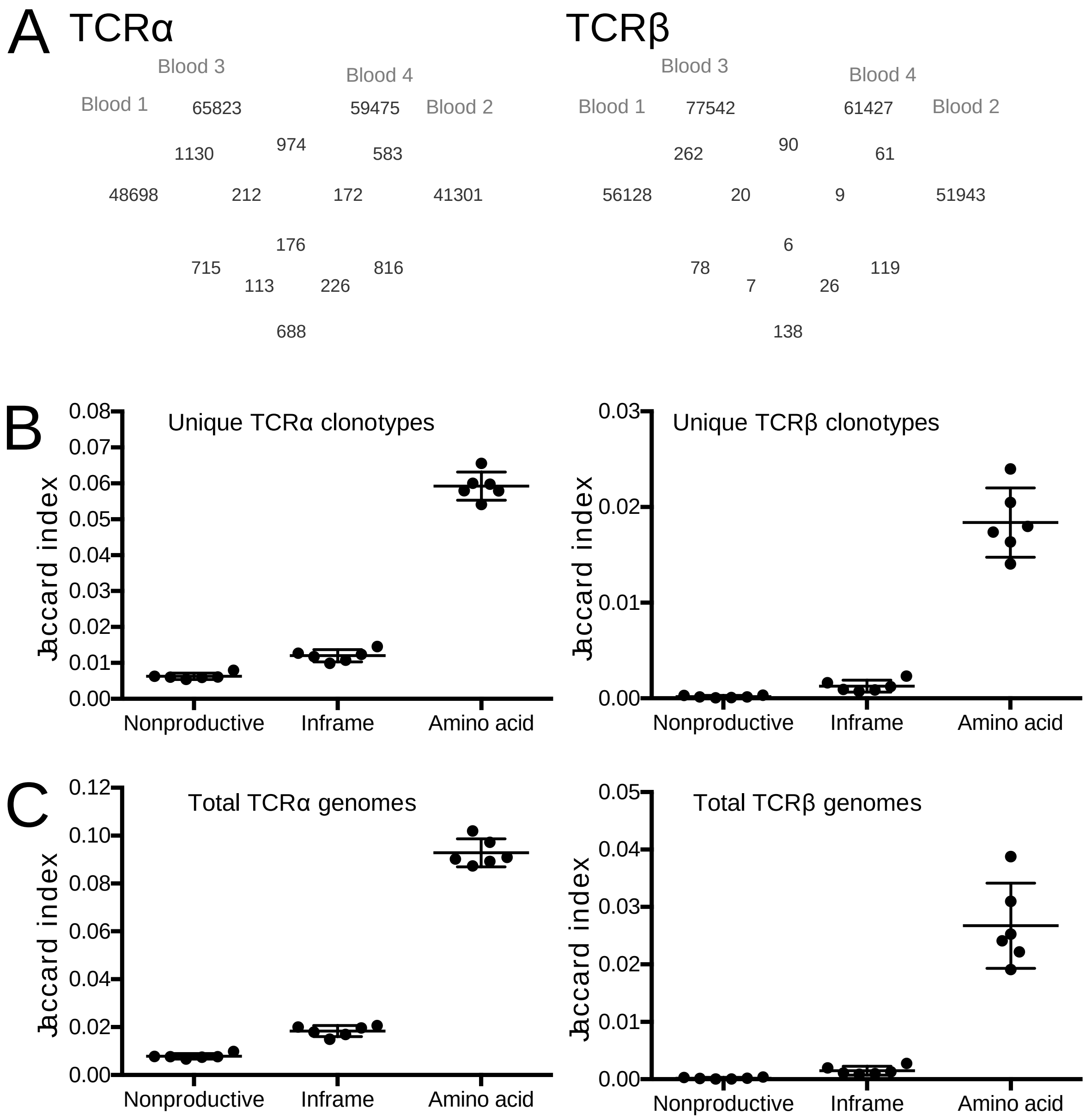

A
TCRα clonotypes
Jaccard index
0.0
0.2
0.4
0.6
0.8
top 1%
top 2%
top 5%
top 10%
top 20%
top 50%
100%
TCRβ clonotypes
Jaccard index
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
top 1%
top 2%
top 5%
top 10%
top 20%
top 50%
100%
B
TCRα clonotypes
Non-templated inserts
0.0
0.5
1.0
1.5
top 1%
top 2%
top 5%
top 10%
top 20%
top 50%
100%
TCRβ clonotypes
Non-templated inserts
0.0
0.5
1.0
1.5
top 1%
top 2%
top 5%
top 10%
top 20%
top 50%
100%

Figure 6


A
TCRα
Blood 3
Blood 4
Blood 1
65823
59475
Blood 2
1130
974
583
48698
212
172
41301
176
715
816
113
226
688
TCRβ
Blood 3
Blood 4
Blood 1
77542
61427
Blood 2
262
90
61
56128
20
9
51943
6
78
119
7
26
138
B
Unique TCRα clonotypes
Jaccard index
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Nonproductive
Inframe
Amino acid
Unique TCRβ clonotypes
Jaccard index
0.03
0.02
0.01
0.00
Nonproductive
Inframe
Amino acid
C
Total TCRα genomes
Jaccard index
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Nonproductive
Inframe
Amino acid
Total TCRβ genomes
Jaccard index
0.05
0.04
0.03
0.02
0.01
0.00
Nonproductive
Inframe
Amino acid