T cell receptor diversity in the human thymus

Vanhanen, Reetta

2016-08

Vanhanen , R , Heikkila , N , Aggarwal , K , Hamm , D , Tarkkila , H , Pätilä , T , Jokiranta , T S , Saramaki , J & Arstila , T P 2016 , ' T cell receptor diversity in the human thymus ' , Molecular Immunology , vol. 76 , pp. 116-122 . https://doi.org/10.1016/j.molimm.2016.07.002

http://hdl.handle.net/10138/224533
https://doi.org/10.1016/j.molimm.2016.07.002

publishedVersion

Molecular Immunology 76 (2016) 116–122

Contents lists available at ScienceDirect

Molecular Immunology

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

T cell receptor diversity in the human thymus

Reetta Vanhanen[a,*], Nelli Heikkilä[a], Kunal Aggarwal[b], David Hamm[c], Heikki Tarkkila[a], Tommi Pätilä[d], T. Sakari Jokiranta[a], Jari Saramäki[b], T. Petteri Arstila[a]

[a] Medicum, Department of Bacteriology and Immunology and Research Programs Unit, Immunobiology, University of Helsinki, 00014 Helsinki, Finland
[b] Department of Computer Science, Aalto University, FI-00076 Aalto, Espoo, Finland
[c] Adaptive Biotechnologies, 1551 Eastlake Ave. E, Seattle, WA, United States
[d] Department of Pediatric Cardiac and Transplantation Surgery, Hospital for Children and Adolescents, Helsinki University Central Hospital, 00290 Helsinki, Finland

ARTICLE INFO

Article history:
Received 30 May 2016
Received in revised form 1 July 2016
Accepted 4 July 2016
Available online 19 July 2016

Keywords:
T cell
Thymus
TCR diversity
Repertoire

ABSTRACT

A diverse T cell receptor (TCR) repertoire is essential for adaptive immune responses and is generated by somatic recombination of TCRα and TCRβ gene segments in the thymus. Previous estimates of the total TCR diversity have studied the circulating mature repertoire, identifying 1 to 3×10^6 unique TCRβ and 0.5×10^6 TCRα sequences. Here we provide the first estimate of the total TCR diversity generated in the human thymus, an organ which in principle can be sampled in its entirety. High-throughput sequencing of samples from four pediatric donors detected up to 10.3×10^6 unique TCRβ sequences and 3.7×10^6 TCRα sequences, the highest directly observed diversity so far for either chain. To obtain an estimate of the total diversity we then used three different estimators, preseq and DivE, which measure the saturation of rarefaction curves, and Chao2, which measures the size of the overlap between samples. Our results provide an estimate of a thymic repertoire consisting of 40 to 70×10^6 unique TCRβ sequences and 60 to 100×10^6 TCRα sequences. The thymic repertoire is thus extremely diverse. Moreover, extrapolation of the data and comparison with earlier estimates of peripheral diversity also suggest that the thymic repertoire is transient, with different clones produced at different times.

1. Introduction

Antigen recognition by T cells is based on the T cell receptor (TCR), a heterodimeric cell surface protein consisting in most cells of α and β chains. To ensure that T cells can react to the great variety of potential pathogens a genetic recombination machinery creates a diverse repertoire of TCRs during the development of T cells in the thymus. The functional genes encoding the antigen-recognizing variable domains are generated by somatic rearrangement of non-contiguous pieces, 52 variable (Vβ), 2 diversity (Dβ) and 13 joining (Jβ) gene segments in the β, and 70–80 Vα and 61 Jα gene segments in the α chain. The TCR chains are further diversified by both deletion of germ-line nucleotides and addition of palindromic P-nucleotides and nontemplated N-nucleotides when the gene segments are joined together. The most diverse area of the TCR is thus the sequence spanning the V(D)J junction, the complementarity determining region 3 (CDR3) which encodes a protein loop directly in contact with the antigenic peptides presented by the MHC (Rudolph et al., 2006).

Theoretical estimates of the potential diversity that can be created by the recombination machinery have ranged up to 10^{15} (Davis and Bjorkman, 1988). Of more interest, however, is the actual diversity of the T cell population, consisting in humans of approximately 5×10^{11} cells (Ganusov and De Boer, 2007). Several factors limit the potential diversity. Although the gene segments can produce thousands of different combinations, it is well established that some rearrangements are favored, while others are used only rarely or not at all (Venturi et al., 2008; Venturi et al., 2011). Junctional diversity is also less than fully stochastic. For example, germ-line-encoded sequences appear at a much higher frequency than chance alone would predict (Robins et al., 2010). Finally, the repertoire is subject to intensive selection, with more than 95% of developing T cells dying in the thymus (Starr et al., 2003). Further selection takes place in the periphery, including also clonal deletion.

An early estimate of the actual diversity, based on sequencing a subset of the repertoire and extrapolation, put the diversity at 10^6 different β chains, 0.5×10^6 α chains, and a minimum of 24×10^6 αβ combinations (Arstila et al., 1999). Another approach, using lim-

Abbreviations: TCR, T cell receptor; DP, double positive; SP, single positive; CDR3, complementarity determining region 3.

* Corresponding author.
E-mail address: reetta.vanhanen@helsinki.fi (R. Vanhanen).

http://dx.doi.org/10.1016/j.molimm.2016.07.002
0161-5890/

iting dilution of T cells and clone-specific PCR, reported that the median frequency of unique TCRβ sequences was 1 in 2.4×10^7 in CD4+cells (Wagner et al., 1998). More recently, the advent of high-throughput sequencing has allowed a larger fraction of the repertoire to be directly analyzed. Warren and colleagues measured 10^6 TCRβ sequences in a healthy blood donor (Warren et al., 2011), while Robins et al. reported a diversity of 3×10^6 different β chains (Robins et al., 2009). A recent study by Qi et al. analyzed replicate blood samples from nine donors (Qi et al., 2014). Although they directly measured only 0.5×10^6 β sequences, by statistical analysis using Chao2 estimator they extrapolated the whole peripheral T cell compartment to contain 100×10^6 TCRβ chains.

To date, the primary TCR repertoire generated in the thymus has not been measured. We provide here the first estimate of human thymic TCR repertoire and also the largest direct measurement of diversity so far obtained.

2. Materials and methods

2.1. Patient samples

Thymic tissue was obtained from a 26-day-old boy, a 4-month-old boy, an 8-month-old boy and an 8-month old girl undergoing corrective cardiac surgery. The tissue is routinely removed for improved exposure during cardiac surgery. From the first donor we also received a blood sample. The pediatric ethics committee of Helsinki University Hospital approved the study, and an informed consent was obtained from the parents of the children. The study was performed in accordance of the Declaration of Helsinki.

2.2. Cell isolation and flow cytometry

Thymocytes were released within 6 h of the thymectomy from the thymus tissue sample by mechanical homogenization. The blood sample was prepared by lysing erythrocytes with brief incubation in sterile aqua. The antibodies used in the experiments were direct fluorochrome conjugates: CD4-APC-Cy7, CD8-PE-Cy7, CD3-PE and CD3-APC (Immunotools, Friesoythe, Germany).

Flow cytometry was performed using the Cyan ADP instrument (Beckman Coulter, USA). Analysis was done with the FlowJo program (FlowJo, LLC, Oregon, USA). Fluorescence compensation settings were optimized by using BD Bioscience CompBeads (Beckton Dickinson, San Jose, CA).

2.3. Genomic DNA extraction and sequencing

Frozen cell samples were processed and analyzed by Adaptive Biotechnologies (Seattle, USA). Genomic DNA extraction was performed according to the manufacturer's instructions (QIAsymphony, Qiagen, Germany). The amount of DNA and the quality of samples were verified before sequencing. The TCR α and β CD3 region was amplified and sequenced from a standardized quantity of DNA using the ImmunoSEQ assay (Adaptive Biotechnologies) (Robins et al., 2009; Sharma et al., 2015). In this assay, a multiplex PCR system was used to amplify the rearranged CDR3β and CDR3α sequences from sample DNA, producing fragments sufficiently long to identify the VDJ region spanning each unique CDR3. Amplicons were sequenced using the Illumina platform. TCRβ V, D and J, and TCRα V and J gene definitions were provided by the IMGT database (www.imgt.org). The assay is quantitative, having used a complete synthetic repertoire of TCRs to establish an amplification baseline and adjust the assay chemistry to correct for primer bias. In addition, barcoded, spiked-in synthetic templates were used to measure the degree of sequencing coverage and residual PCR bias. This information was used for further PCR bias correction and to estimate the abundance of sequenceable templates in each sample. The resulting data was filtered and clustered using both the relative frequency ratio between similar clones and a modified nearest-neighbor algorithm, to merge closely related sequences and remove both PCR and sequencing errors. Data was analyzed using the ImmunoSEQ analyzer toolset. The final output of this analysis is an estimate of the actual number of cells of each clonotype. The number of observed clonotypes directly yields a lower bound of T cell diversity.

2.4. Extrapolation of total TCR diversity by rarefaction curves

We first constructed rarefaction curves from the sequence data (number of unique TCRβ or α sequences as a function of observed cells), also called complexity curves (Daley and Smith, 2013). The curves were constructed by generating a vector of TCR sequences with as many entries per sequence as observed in the data, randomly reordering the elements of the vector, and then reading the elements one by one while keeping track of the number of unique sequences. We then extrapolated the resulting curves to $n = 1.3 \times 10^9$ cells with two software packages, preseq (Daley and Smith, 2013) and DivE (Laydon et al., 2015; Laydon et al., 2014).

The C++ package preseq (Daley and Smith, 2013) (http://smithlab.usc.edu/software/librarycomplexity/) uses a method of extrapolation that is based on a nonparametric Bayesian model. On the basis of the rarefaction curve where the counts are treated as Poisson random variables, preseq computes a power-series formula that estimates how many times each sequence would be observed in a similar experiment of the same size. The power series is then used to extrapolate the observations, with the help of a technique called rational function approximation that helps to make the series converge when extrapolating to larger samples. Preseq computes 95% confidence intervals using bootstrapping.

The R package DivE (Laydon et al., 2014) (https://cran.r-project.org/web/packages/DivE/) extrapolates rarefaction curves by fitting 58 different functions to the curves and their subsamples. DivE then scores each function and computes the geometric average of the extrapolations of the 5 best-scoring functions. The scores given to functions are based on four criteria: discrepancy (how good the fit is), accuracy (how well the full observed diversity is predicted from subsamples), similarity (how well the functions fit to the whole data and a subsample match), and plausibility (observed diversity needs to grow or plateau). DivE does not produce confidence intervals for individual functions; however, the spread of the 5 best-scoring functions can be considered as indicative of accuracy.

2.5. Extrapolation of total TCR diversity by incidence data

As a complementary approach we used the nonparametric estimator Chao 2 (Chao, 1987), commonly used to estimate species diversity in ecological and microbiological studies. Chao 2 considers each unique TCR sequence as a species. It estimates the total species diversity based on incidence data, a binary matrix where columns represent independent samples of the population and rows represent presence of species in these samples. High column overlap then means that most of the diversity has been sampled; low overlap means that the number of unseen species is expected to increase upon further sampling. The R package *fossil* was used for implementing Chao 2 (Vavrek, 2011). To verify the robustness of the Chao 2 estimates, we bootstrapped by randomly picking half the data from each dataset, building the incidence matrix, running Chao 2, and recording the estimate. This was repeated 100 times.

Table 1
Sequencing results of TCRβ genes in the samples 1–4.

	Age and gender	Cell count	Total reads	In-frame	In-frame%	Unique in-frame
1A	26 days, male	12 million	46 355 113	36 097 248	77.9	3 775 569
1B		12 million	45 284 758	35 323 155	78.0	3 831 886
1C		12 million	50 553 303	39 299 132	77.7	3 995 054
2A	8 months, male	4 million	11 558 445	8 803 558	76.2	1 176 610
2B		4 million	21 901 828	16 821 051	76.8	1 686 993
3	4 months, male	4 million	23 581 729	20 068 112	85.1	1 568 528
4	8 months, female	4 million	11 159 872	9 024 467	80.9	1 462 150

Table 2
Sequencing results of TCRα genes in the samples 2–4.

Sample	Age and gender	Cell count	Total reads	In-frame	In-frame%	Unique in-frame
2A	8 months, male	4 million	45 335 572	14 585 147	32.2	2 398 350
2B		4 million	27 172 083	8 694 216	32.0	1 542 907
3	8 months, male	4 million	30 309 225	9 640 760	31.8	1 684 426
4	4 months, female	4 million	36 762 724	11 754 417	32.0	2 125 962

3. Results

3.1. Sample preparation and sequencing of TCR genes

Our study included four pediatric thymus samples, obtained from otherwise healthy children undergoing corrective cardiac surgery. Thymus sample 1 was homogenized in its entirety and contained 1.3×10^9 thymocytes. Genomic DNA was extracted from the total amount of cells and three samples, each the equivalent of 12×10^6, cells were analyzed. Of the three other thymuses (thymuses 2–4), samples of 10×10^6 thymocytes were isolated and the gDNA equivalent of 4×10^6 cells analyzed. From thymus 2, two separate samples of 10×10^6 thymocytes were extracted and analyzed.

To ascertain that the thymuses were normal we used flow cytometry to analyze the main thymocyte subsets. All four samples showed a normal distribution of CD4/CD8 double negative (mean 1.6%, range 0.7–2.1%), double-positive (79.5%, range 76.5–79.5%), CD4+ single-positive (11.8%, range 10.2–13.4%), and CD8+ single-positive cells (8.6%, range 7.0–10.8%) (Fig. 1). As expected, most double-negative thymocytes were CD3−, double-positive cells showed a distribution ranging from CD3− to CD3high, and the single-positive subsets were almost completely CD3high.

The complementarity determining region 3 (CDR3) sequences of TCRβ were then amplified and sequenced from all four thymus samples. On the average, 47.4×10^6 TCRβ reads were obtained from the three replicates of thymus 1, and 17.1×10^6 reads from thymuses 2–4 (Table 1). Approximately 80% of the sequences were productive. The number of unique productive TCRβ sequences found ranged from 1.5×10^6 in thymus 4 to 10.3×10^6 in thymus 1 when all three replicates were combined, the highest TCR diversity so far directly measured.

The complementarity determining region 3 of TCRα/δ was sequenced from thymus samples 2–4. Because some T cells express α chains utilizing Vδ or Jδ gene segments (Chien et al., 1987; Satyanarayana et al., 1988), with our approach it was not possible to reliably separate TCRα and TCRδ loci, and a small fraction of the sequences were therefore likely to be derived from γδ T cells. In all samples, less than 5% of the sequences utilized either Vδ or Jδ genes. On the average, 34.9×10^6 reads were obtained, 32% of which were productive (Table 2). This low frequency of productive rearrangements is most likely due to the lack of allelic exclusion in the TCRα locus. The number of unique productive sequences ranged from 1.7×10^6 in thymus 3 to 3.7×10^6 in replicates of thymus 2 combined, again the highest directly measured TCRα diversity to date.

Table 3
TCR diversity estimated by Preseq.

Sample	Estimate	95% confidence interval
TCRβ		
1	73.1×10^6	43.3 to 129.9×10^6
2	37.6×10^6	26.6 to 54.0×10^6
3	18.7×10^6	10.2 to 35.5×10^6
4	13.9×10^6	11.5 to 16.8×10^6
TCRα		
2	83.4×10^6	49.8 to 146.8×10^6
4	37.7×10^6	25.5 to 57.9×10^6

3.2. Estimation of the total TCRβ diversity in thymus 1

Despite the 10.3×10^6 unique sequences obtained from thymus 1, the number of sequences shared by any two of the three replicates was only 0.5×10^6. The Jaccard index, obtained by dividing the number of shared sequences by the combined size of the samples was only 0.07, indicating that only a fraction of the total diversity had been sampled. We therefore used three different approaches to estimate the size of the full repertoire by extrapolation. Two were based on measuring the saturation of rarefaction curves and one on the size of the overlap between samples.

Rarefaction curves measure the number of observed unique TCRβ sequences as a function of the number of sequenced cells. Rarefaction curves are increasing functions that saturate when the complete diversity has been sampled. Since our sample sizes were far below saturation the curves needed to be extrapolated to saturation. We first used the preseq software package developed by Daley and Smith that extrapolates the curve with a power series using the Rational Function Approximation method (see Methods) (Daley and Smith, 2013). Using preseq produced an estimate of 73.1×10^6 unique TCRβ sequences in the whole thymus, with 95% confidence interval of 43.3 to 129.9×10^6 (Table 3, Fig. 2). Another method of extrapolation from rarefaction curves, DivE was then used (Laydon et al., 2015; Laydon et al., 2014). DivE fits a large number of functions to the curves and scores each with 4 criteria that measure the accuracy and consistency of fits to data and subsamples (see Methods). DivE then uses the geometric mean of the five best models as the final estimate. Extrapolating by DivE indicated a total of 42.5×10^6 unique TCRβ sequences in thymus 1 (Table 4, Fig. 2).

The third estimator, Chao2, estimates the total diversity based on incidence data, more specifically the overlap of TCRβ sequence incidence (presence/absence) vectors between different data sets or samples (Chao, 1987). Chao2 estimated a total diversity of

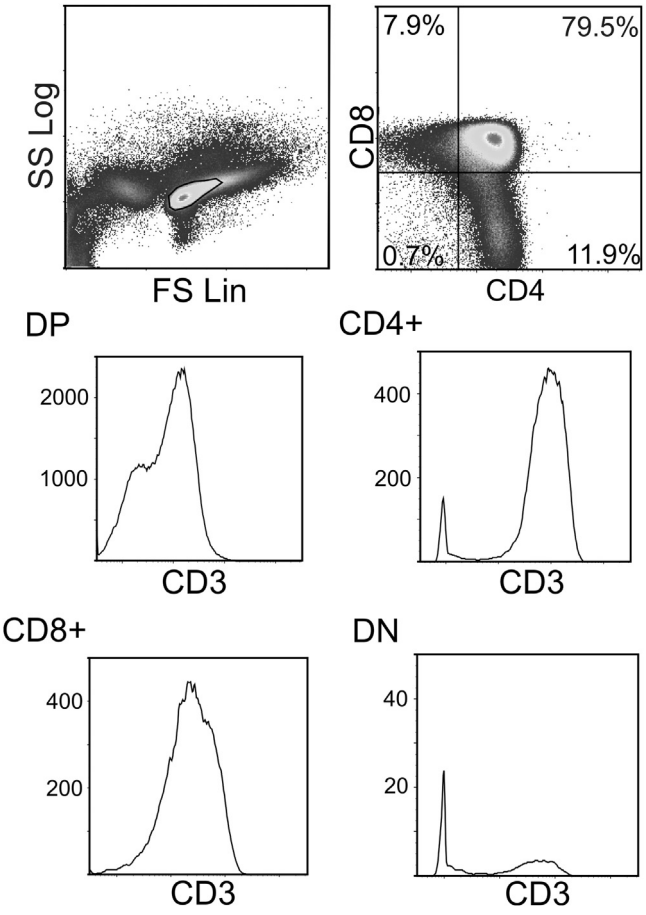

Fig. 1. Flow cytometric analysis of thymus 1.
Thymocyte gating and distribution of main thymocyte subsets. The data is shown on logarithmic scale, except for forward scatter (FS). The frequency of each main subset is indicated in the figure. CD3 exression profiles of the main subsets are shown in the histograms. DN: CD4/CD8 double-negative, DP: CD4/CD8 double-positive.

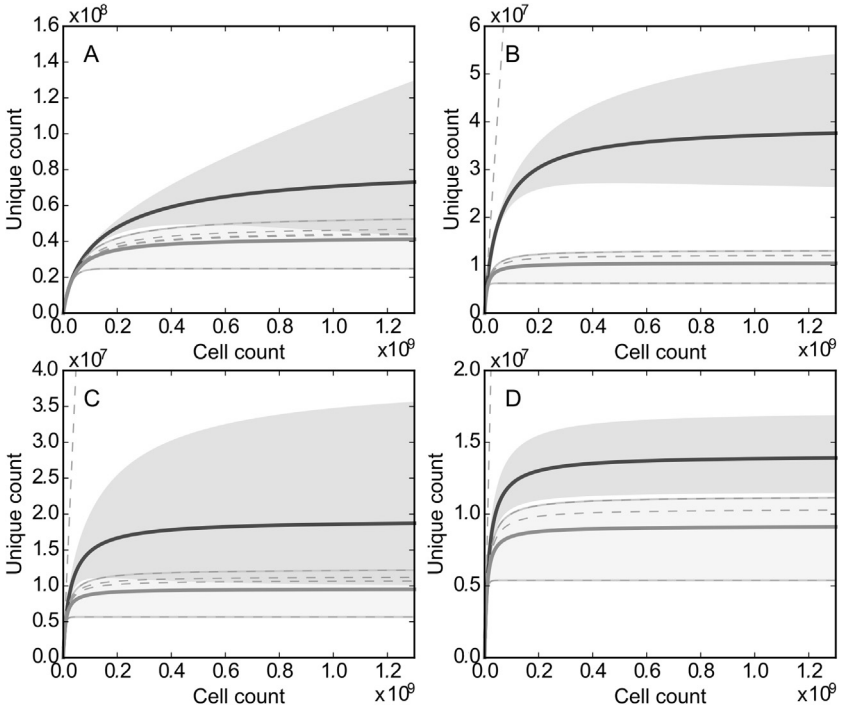

Fig. 2. Extrapolation of rarefaction curves for TCRβ diversity.
Extrapolation by preseq is shown solid dark line, with 95% confidence interval indicated by shading shading. The 5 best-scoring extrapolations obtained by DivE are shown as dashed lines, and their geometric mean as solid grey line. A-D indicate thymuses 1–4, respectively. The outlier DivE extrapolations in B) and C) are excluded from the mean.

Table 4
TCR diversity estimated by DivE ($\times 10^6$).

Sample	Model 1	Model 2	Model 3	Model 4	Model 5	Final Est
TCRβ						
1	55.1	**24.8**	45.4	45.9	48.6	**42.5**
2	13.2	6.3	**12.2**	12.2	(783,813.9)	**10.5**
3	11.3	5.7	10.8	**12.4**	(776,228.4)	**9.6**
4	11.2	5.4	**10.4**	11.2	(782,072.4)	**9.2**
TCRα						
2	31.9	13.3	**60.4**	24.1	3.2	**28.8**
3	24.3	**10.3**	19.0	19.5	24.3	**18.6**
4	28.1	12.1	**22.1**	28.5	28.1	**22.7**

The numbers in bold indicate the estimate of the best-scoring model. The final estimate is the geometric mean of the 5 best models.

Table 5
TCR diversity estimated by Chao2.

Sample	Estimate	Bootstrap estimate
TCRβ		
1	46.4×10^6	43.0×10^6
2	30.0×10^6	28.1×10^6
3	10.1×10^6	9.7×10^6
4	9.6×10^6	9.0×10^6
TCRα		
2	103×10^6	62.3×10^6
3	16.9×10^6	14.0×10^6
4	19.5×10^6	16.3×10^6

46.4×10^6 unique sequences (Table 5). To verify the robustness of the Chao 2 estimate, we bootstrapped by randomly picking half the data from each three data sets and computing the estimate, repeating the procedure 100 times. This yielded slightly lower values than the original estimates ($43.0 \times 10^6 \pm 0.1 \times 10^6$).

A small volume of blood was available from the donor of thymus 1. Sequencing of 3.4×10^6 total TCRβ reads produced 86 000 unique productive sequences, 4000 of which were shared with the thymus repertoire.

3.3. Estimation of the total TCRβ diversity in thymuses 2–4

The three methods of extrapolation were then applied to thymus samples 2–4. Estimates based on preseq ranged from 13.9×10^6 to 37.6×10^6, while DivE indicated 9.2 to 10.5×10^6 unique sequences. For thymuses 2 and 3 one of the five best DivE models failed to saturate at all, and the final estimate was based on only four models (Fig. 2). The overlap-based estimator Chao2 was used by comparing the two replicates of thymus 2 and randomly divided halves of data sets from thymuses 3 and 4. Extrapolation by Chao2 yielded an estimate of 30.0×10^6, 10.0×10^6 and 9.6×10^6 sequences in thymuses 2–4, respectively. Again, the bootstrap estimates computed using half the available data were somewhat smaller (28.1×10^6, 9.7×10^6, and 9.6×10^6) (Table 5).

3.4. Estimation of the total TCRα diversity in thymuses 2–4

Analysis of TCRα/δ repertoire in thymuses 2–4 using the three estimators produced estimates of total diversity, which were in every case higher than the corresponding estimate of TCRβ diversity. Preseq extrapolated to 83.4×10^6 (95% CI 49.8×10^6 to 146.8×10^6) sequences in thymus 2 and 37.7×10^6 (25.5×10^6 to 57.9×10^6) in thymus 4 (Table 3). The estimator failed for thymus 3. Using DivE extrapolated to 28.8×10^6, 18.6×10^6 and 22.7×10^6 unique TCRα sequences in thymuses 2–4, respectively (Table 4, Fig. 3). Finally, estimates based on Chao2 were 103×10^6, 16.9×10^6 and 19.5×10^6 TCRα sequences in thymuses 2–4 (Table 5), respectively. Bootstrap estimates obtained by running Chao2 on half the data, as described for TCRα, were 62.3×10^6, 14.0×10^6 and 16.3×10^6. Although to some extent divergent, these results show that extrapolated TCRα diversity is consistently higher than the corresponding TCRβ diversity.

4. Discussion

The present report provides the first estimate of the human thymic T cell diversity and the largest directly measured repertoire so far. We obtained 10.3×10^6 different TCRβ sequences from thymus 1 and 3.7×10^6 α sequences from thymus 2, while statistical extrapolation indicated that the total thymic TCRβ diversity is 40 to 70×10^6 and α diversity is approximately twice as high. An earlier study of peripheral repertoire showed that each β chain in the circulating repertoire can pair with at least 24 different α chains (Arstila et al., 1999). Given the extensive cell deletion in the thymus, this ratio is unlikely to be quite as high in the thymocytes. This uncertainty concerning α to β pairing in thymus remains an obstacle in providing an exact numerical value for the number of αβ heterodimers. Nevertheless, our data put the total thymic αβ TCR repertoire in the hundreds of millions different receptors.

Unlike estimates of peripheral repertoire, which by necessity are based only sampling only a small fraction of the whole lymphoid compartment, the primary repertoire in the thymus can in principle be sampled and analyzed in its entirety. However, the extremely high diversity and limitations in current sequencing technology still oblige to use extrapolation. Our estimates are based on three different approaches, preseq, DivE and Chao2, all with potential confounding factors. First, DivE uses empirical, nonstandard, ad-hoc criteria to score fitted models; the set of models used has been a subjective choice. Further, for rarefaction curves that display low curvature, i.e. that are very far from the saturation point, both preseq and DivE may produce incorrect estimates. However, in this case, other methods will fail too, as linearly growing rarefaction curves simply mean that the diversity has not been sampled well enough. This is also the case for Chao 2: Chao et al. reports that the estimator may break down when the overlap between samples is low (Chao, 1987). While in our case the rarefaction curves do display some curvature, it is clear from Fig. 2 that there is uncertainty because the estimates of individual DivE models cover a range of values. However, the three methods produced generally convergent outcomes, both between estimators and between samples. It is also notable that with increasing sample size the extrapolations produced by the estimators were also consistently higher. It is thus likely, with a reasonable degree of confidence that our higher estimates are not very far from the truth. This holds especially for β chain, which was studied in more detail. The numerical estimate of α diversity is less solid, but the diversity does seem higher than β diversity.

Although α chain diversity has been studied very little, the existing data suggest that in the peripheral repertoire α diversity is lower than in the β chain (Arstila et al., 1999), whereas our data indicate the opposite in thymocytes. This reversed ratio is even more notable when the timing of TCR rearrangements in the two loci is taken into account. Because TCRβ locus is rearranged before α, a substantial fraction of thymocytes in our analysis expresses only the β chain, but even so the α diversity exceeds β. The effects of the main thymic maturation pathway then become apparent in the pruning of the α repertoire. After successful expression of the TCR β chain, and before α rearrangement, the cells proliferate extensively (Dik et al., 2005). Therefore, at the time of α rearrangement each β chain will be expressed by a large number of cells and will end up paired with a large number of different α chains (Padovan et al., 1993). The bulk of thymic selection and thymocyte death takes place after α rearrangement and the surface expression

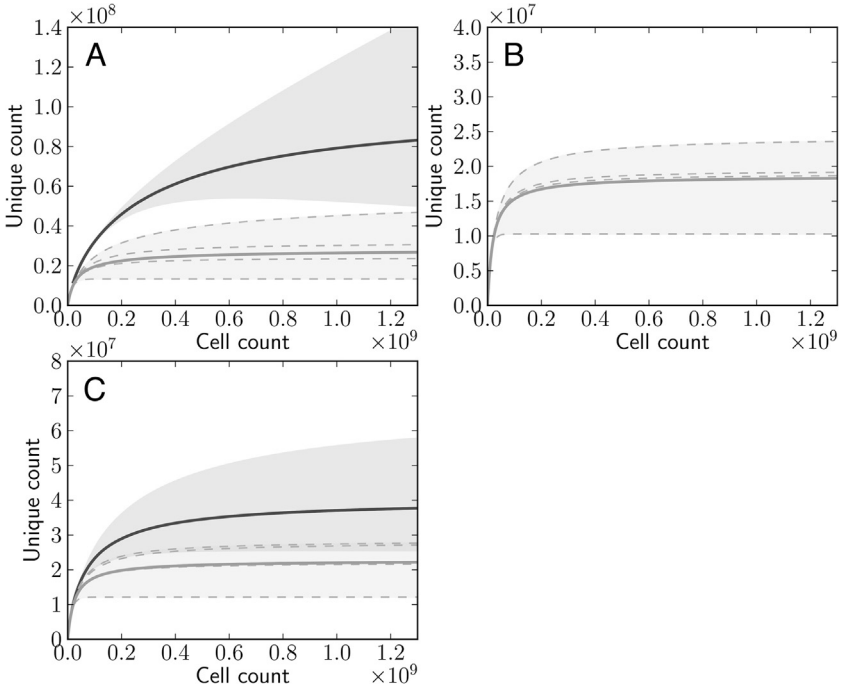

Fig. 3. Extrapolation of rarefaction curves for TCRα diversity.
Extrapolation by preseq is shown as solid dark line, with 95% confidence interval indicated by shading shading. The 5 best-scoring extrapolations obtained by DivE are shown as dashed lines, and their geometric mean as solid grey line. A-C indicate thymuses 2–4, respectively. Preseq failed to produce an estimate for data set 3 (panel B).

of αβ TCR. The difference between thymic and peripheral α diversity thus reflects the difference between rearranged and selected α repertoire. These considerations also suggest that although so far repertoire studies have predominantly analyzed β chain, α repertoire might be a more sensitive indicator of clonal changes.

Our estimate of 40 to 70 × 10^6 thymic β chains was also smaller than the biggest estimate of peripheral β repertoire at 100 × 10^6, a difference likely to be even bigger, when the effects of thymocyte deletion are taken into account. This emphasizes the transient and dynamic nature of the thymocyte population as the source of peripheral T cell compartment. It also implies that the clonal composition of thymic repertoire changes constantly, and different clones are produced at different times. It is thus important to note that our estimates represent one given time point, a consideration more important for thymic repertoire than for peripheral, since the former is likely to be more transient, given the continuous turnover of thymocytes caused by death and egress. In donor 1, of whom a small volume of blood was available, 4.7% of the peripheral β sequences were also found in the thymus. This suggests some degree of continuity in thymic clonal production, although we cannot exclude recirculation of peripheral clones into thymus as contributing to the overlap. Nevertheless, since less than 5% of the circulating clones were also found in the thymus, although close to 20% of the extrapolated total thymus diversity was sequenced, most of the peripheral repertoire was distinct of the thymic repertoire at the time of sampling. Although obviously very preliminary, this observation further indicates that the clonal composition in thymus is variable.

Taken together, we provide the first ever estimate of the size of the human thymic TCR repertoire and the highest directly measured diversity to date. Our data indicate that thymic repertoire is extremely diverse, but also unstable, with different clones produced at different times. The extensive clonal deletion of thymocytes is apparent in the loss of TCRα diversity, when thymus and periphery are compared.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Tamás Bazsinka for technical assistance. This work was funded by the Doctoral Programme in Biomedicine, University of Helsinki, Finnish Medical Foundation, and University of Helsinki Research funds. R.V received a Young Investigator's Award from Adaptive Biotechnologies.

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.
Chao, A., 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783–791.
Chien, Y.H., Iwashima, M., Kaplan, K.B., Elliott, J.F., Davis, M.M., 1987. A new T-cell receptor gene located within the alpha locus and expressed early in T-cell differentiation. Nature 327, 677–682.
Daley, T., Smith, A.D., 2013. Predicting the molecular complexity of sequencing libraries. Nat. Methods 10, 325–327.
Davis, M.M., Bjorkman, P.J., 1988. T-cell antigen receptor genes and T-cell recognition. Nature 334, 395–402.
Dik, W.A., Pike-Overzet, K., Weerkamp, F., de Ridder, D., de Haas, E.F., Baert, M.R., van der Spek, P., Koster, E.E., Reinders, M.J., van Dongen, J.J., Langerak, A.W., Staal, F.J., 2005. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J. Exp. Med. 201, 1715–1723.
Ganusov, V.V., De Boer, R.J., 2007. Do most lymphocytes in humans really reside in the gut? Trends Immunol. 28, 514–518.
Laydon, D.J., Melamed, A., Sim, A., Gillet, N.a., Sim, K., Darko, S., Kroll, J.S., Douek, D.C., Price, D.a., Bangham, C.R.M., Asquith, B., 2014. Quantification of HTLV-1 clonality and TCR diversity. PLoS Comput. Biol. 10, 1–13.
Laydon, D.J., Bangham, C.R., Asquith, B., 2015. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Philos. Trans. R. Soc. Lond. B: Biol. Sci. 370.
Padovan, E., Casorati, G., Dellabona, P., Meyer, S., Brockhaus, M., Lanzavecchia, A., 1993. Expression of two T cell receptor alpha chains: dual receptor T cells. Science 262, 422–424.

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.

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.

Robins, H.S., Srivastava, S.K., Campregher, P.V., Turtle, C.J., Andriesen, J., Riddell, S.R., Carlson, C.S., Warren, E.H., 2010. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci. Transl. Med. 2, 47ra64.

Rudolph, M.G., Stanfield, R.L., Wilson, I.A., 2006. How TCRs bind MHCs, peptides, and coreceptors. Annu. Rev. Immunol. 24, 419–466.

Satyanarayana, K., Hata, S., Devlin, P., Roncarolo, M.G., De Vries, J.E., Spits, H., Strominger, J.L., Krangel, M.S., 1988. Genomic organization of the human T-cell antigen-receptor alpha/delta locus. Proc. Natl. Acad. Sci. U. S. A. 85, 8166–8170.

Sharma, P.K., Wong, E.B., Napier, R.J., Bishai, W.R., Ndung'u, T., Kasprowicz, V.O., Lewinsohn, D.A., Lewinsohn, D.M., Gold, M.C., 2015. High expression of CD26 accurately identifies human bacteria-reactive MR1-restricted MAIT cells. Immunology 145, 443–453.

Starr, T.K., Jameson, S.C., Hogquist, K.A., 2003. Positive and negative selection of T cells. Annu. Rev. Immunol. 21, 139–176.

Vavrek, M.J., 2011. Fossil: palaeoecological and palaeogeographical analysis tools. Palaeontol. Electronica 14.

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., 2008. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. J. Immunol. 181, 7853–7862.

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.

Wagner, U.G., Koetz, K., Weyand, C.M., Goronzy, J.J., 1998. Perturbation of the T cell repertoire in rheumatoid arthritis. Proc. Natl. Acad. Sci. U. S. A. 95, 14447–14452.

Warren, R.L., Freeman, J.D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J.R., Holt, R.A., 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 21, 790–797.