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.