ESTIMATING THE NUMBER OF HIV-SPECIFIC T-CELLS IN HEALTHY DONORS USING HIGH-THROUGHPUT SEQUENCING PROFILES OF T-CELL RECEPTOR REPERTOIRES

Eliseev AV, Fedorova AD, Lebedin MY, Chudakov DM, Shugay M

Pirogov Russian National Research Medical University, Moscow, Russia

In-depth study of mechanisms of immune response to the human immunodeficiency virus (HIV) is critical for better understanding of how immunodeficiency develops in patients with HIV, as well as for designing effective immunotherapy strategies and vaccines against the virus. In this work we analyze sequencing profiles of T-cell receptor repertoires previously obtained from healthy donors (601 Americans and 65 Russians) to estimate the population frequency of HIV-specific naive T-cells. We demonstrate that frequencies of T-cells recognizing different HIV epitopes vary considerably across the population (F-statistic = 2007, p < 10^{-10}, ANOVA). Although the frequency of T-lymphocytes recognizing a particular epitope does not change significantly between the individuals, it still largely depends on the presence of certain HLA alleles (p < 0.01, post-hoc Tukey’s test), cytomegalovirus infection (F = 61, p = 7 \times 10^{-15}, ANOVA), and age (Pearson correlation coefficient ranging from –0.53 to –0.14 in different groups).

Keywords: HIV, T-cell receptor, HLA, CMV, high-throughput sequencing

Funding: this work was supported by the Russian Science Foundation (Grant No. 17-15-01495).

Correspondence should be addressed: Mikhail Shugay ul. Miklukho-Maklaya, d. 16/10, Moscow, Russia, 117997; mikhail.shugay@gmail.com

Received: 23.10.2017 Accepted: 28.10.2017

The highly mutable human immunodeficiency virus (HIV) easily evades the immune system [1]. Still, there is hope for anti-HIV immunotherapy, considering the variety of immunogenic HIV epitopes [2] and protective human leukocyte antigen (HLA) alleles [3], as well as the phenomenon of the so-called elite controllers [4]. Although attempts to develop an HIV vaccine have not paid off yet, the accumulated evidence suggests that the T-cell therapy may potentially be as effective as the conventional antiretroviral treatment [5].

In this work we draw on the assumption that the proportion of antigen-specific T-cells occurring in the naive T-lymphocyte population determines the magnitude of the immune response [6]. Knowing which HIV epitopes tend to be readily recognized by the immune system of a person who carries a particular set of HLA alleles will help to elucidate mechanisms of immune protection against HIV and finally make headway in the development of an HIV vaccine.

The ability of T-cells to recognize foreign antigens is encoded by alpha- and beta-chain genes of T-cell receptors (TCR). Their diversity is incredible: the number of unique sequences (variants) in a person’s beta chain is estimated to be over 10^9, while the total number of TCR variants generated in the thymus...
gland of each member of the population is almost infinite [7]. Massively parallel sequencing of immune repertoires (RepSeq) has evolved to simultaneously produce millions of TCR reads per studied sample, e.g., of peripheral blood mononuclear cells [8]. Currently existing methods of T-cell sorting, especially those based on MHC multimer staining [9], yield a wealth of information about antigen-specific TCR. In this light, RepSeq can be conveniently used to analyze individual TCR repertoires. For example, data generated by RepSeq can be further annotated in silico and the number of epitope-specific T-cells can be estimated using the regularly updated VDJdb repository of TCR sequences with known antigen specificity [10].

That said, it is almost impossible to accurately quantify antigen-specific T-cells in the naive T-cell population using standard techniques, such as flow cytometry. Because the population of T-cells that recognize a particular epitope is often very small (<1%) [11], magnetic bead enrichment may be needed [12], which, unfortunately, can distort the results. In contrast, RepSeq reliably reports T-cells with frequencies as low as 0.001% [13].

Having a large dataset of TCR sequences at our disposal obtained from 65 Russian and 601 American donors and another dataset of 1,688 TCR with known epitope specificity (see Methods), we have attempted to study the frequency of HIV-specific T-cells in the population. The following hypotheses have been tested:

1) frequencies of epitope-specific T-cells in the TCR repertoires of healthy individuals vary considerably depending on the epitope;
2) cytomegalovirus (CMV) infection in the individual affects the proportion of HIV-specific T-cells in his T-cell repertoire;
3) the number of HIV-specific T-cells depends on the presence of specific HLA alleles in the individual;
4) the number of HIV-specific T-cells depends on the individual's age and sex.

METHODS

We analyzed the datasets of sequenced TCR beta chains obtained by Emerson et al. [14] and Britanova et al. [15]. We...
did not use all of the sequenced data obtained by Emerson, selecting the TCR repertoires of only those donors whose HLA haplotype had been identified and CMV status was known — a total of 601 samples. We also filtered out umbilical cord blood TCR from Britanova et al. study’s sample, saving for the analysis only the repertoires of 65 healthy adults. Data preprocessing and segment mapping for sequences borrowed from [15] were performed with MiGEC [16] and MiTCR [17] software tools. Segments from [14] were additionally mapped, V- and J-segment genes were identified and sequencing errors were corrected using MiXCR [18]. Data were cleaned of non-functional sequences containing stop-codons or frameshifts using VDJtools [19].

Annotation, i.e. prediction of HIV-specific TCR, was done using VDJtools/VDJdb-standalone [10]. VDJdb was searched for HIV-specific TCR; epitopes represented in the database by less than 10 TCR variants were excluded from the analysis. A RepSeq TCR was counted as specific to a particular epitope if the amino acid sequence of the epitope’s hyper variable CDR3 region differed by no more than 1 substitution from the corresponding TCR sequence stored in VDJdb. This approach yields a substantially larger set of annotated TCR, with only a tiny percent of erroneous annotations, as shown in [10].

Statistical analysis was done with R. The following statistical algorithms were used: ANOVA, Tukey’s post hoc test and correlation analysis. Values for the F-statistic, Spearman’s rank correlation and Student’s p are provided in the Results section.

**RESULTS**

Frequency estimates obtained by flow cytometry for HIV-specific TCR convincingly demonstrate that the proportion of specific T-cells in the naive (intact) repertoire varies considerably, differing by 1 or 2 orders of magnitude between the epitopes, while remaining fairly stable between different individuals [12]. Analysis of high-throughput TCR sequencing data conducted in the course of our study (Fig. 1) supports these observations: frequencies of HIV-specific TCR have been found to be highly epitope-dependent ($F = 2007, p < 10^{-100}$, ANOVA), which, however, bears no connection to the presenting HLA allele ($F = 0.03, p = 0.86$, ANOVA). Importantly, there is a significant discrepancy in the estimates for Emerson’s and Britanova et al. study’s datasets ($F = 1690, p < 10^{-100}$; average frequency of HIV-specific TCR is higher for Emerson et al. study’s data), which can be explained by different structures of TCR libraries and techniques used for their preparation. Emerson et al. worked with DNA samples employing multiplex PCR, while Britanova et al. used RNA samples, 5’ RACE and molecular barcoding [20]. Skipping the details, we will, however, emphasize that molecular barcoding ensures more accurate quantification of TCR in the sample [15].

In the study by Emerson et al. the donors were divided into two cohorts based on their serologic status, i.e. on the presence or absence of CMV infection. With sequencing data at our disposal, we seized this opportunity to evaluate...
the impact of CMV infection on the frequency of HIV-specific TCR in donors’ repertoires. As shown in Fig. 2, the frequency of HIV-specific TCR was significantly higher for CMV-negative individuals regardless of the HIV epitope (F = 495, p < 10^{-100}, ANOVA). Of note, if TCR were not grouped based on the epitope they recognize, i.e. if the epitope-related difference in HIV-specific TCR frequencies was ignored, the result would be far less significant (F = 61, p = 7×10^{-16}, ANOVA).

Information about the HLA haplotypes of the donors provided by Emerson et al. was used to estimate the number of HIV-specific TCR considering that the donor may have provided by Emerson et al. was used to estimate the number of T-cells capable of recognizing these epitopes.

As shown in Fig. 3, the largest proportion of HIV-specific TCR is observed for putatively protective B27, B57 and B51 HLA alleles [3]. For these 3 alleles the number of HIV-specific TCR is significantly higher than for 5 other alleles (p < 0.01, Tukey’s post hoc test), except for the differences between alleles B51 and B08. It should be noted that we had to recruit a relatively small number of alleles for out study because there were no known HIV epitopes for other alleles in VDJdb.

Age-related changes in the structure of the T-cell repertoire were described in a number of previously published works [21] reporting the reduction of the observed repertoire diversity due to clonal expansions caused by chronic infections. Impoverished diversity results in the decreased proportion of T-cells (including the HIV-specific T-cells) capable of recognizing previously unencountered pathogens (Fig. 4):
Fig. 4. Changes in the total frequency of HIV-specific TCR in donors of different sex and age. Data from Emerson et al. are distributed into groups based on the CMV status of the participants. The graph also shows results of linear modeling for HIV-specific TCR dependence on age for males and females.

References


