Comparative analysis of V-D-J rearrangement junction sequences that encode T-cell receptors recognizing similar and distinct antigens.

Aleksey V. Eliseev, Dmitry M. Chudakov, Mikhail Shugay
Pirogov Russian National Research Medical University, Ostrovityanova str. 1, Moscow, Russia, 117997,
aleksey.yelis@gmail.com

T-cell receptor (TCR) sequences encode the antigen specificity of the cellular branch of the adaptive immunity and its ability to mount an effective immune response against novel and previously encountered pathogens. Recent advances in high-throughput sequencing techniques allows profiling millions of TCRs from the sample of interest, producing a snapshot of the T-cell repertoire [1]. The immense diversity of TCR sequences (more than 10^12 unique variants of TCR beta chain) together with the extremely small number of annotated sequences in public databases (around 10^3 corresponding entries in GenBank, many lacking proper annotation) makes it very hard to interpret the T-cell repertoire sequencing data in terms of antigen specificities. Thus the analysis of TCR repertoires is currently limited to a set of summary statistics [2, 3], which is one of the major problems on the way of utilizing the full potential of immune repertoire sequencing technology both in the area of basic research and in personalized medicine.

To address the problem of functional annotation of TCR repertoires, we have manually collected more than 2300 previously published high-quality human TCR beta chain sequences that recognize specific epitopes from pathogens that are both common (Cytomegalovirus, Epstein-Barr Virus) and relatively rare (HIV-1, Hepatitis C virus). Comparative analysis of TCR sequences reveals that there is a higher degree of sequence similarity between TCRs specific to the same epitope than between those specific to different ones. Moreover, the analysis of the graph of TCR sequences constructed under a specific hamming distance threshold reveals a clear co-clustering of TCRs specific to the same epitope. These findings suggest that a properly designed TCR sequence alignment algorithm can be used to extend the annotation from the relatively small set of manually curated TCR sequences to the scale of a typical peripheral blood repertoire sample which contains around 1mln unique TCR sequences [4].

Finally, we demonstrate the utility of our dataset for the inference of amino acid motifs from the CDR3 regions of TCR sequences specific to certain epitopes. Notably, these motifs lie predominantly in the N-region of CDR3 that is generated from random nucleotide insertions during V-D-J recombination process and become even more evident when correcting for an

extremely strong germline bias coming from V and J segments of TCR sequence. By comparing our motifs to the available TCR:peptide:MHC structural data, we show that the identified amino acid motifs correspond exactly to the positions in the CDR3 region that are in a direct contact with antigen residues.

This work was supported by Russian Science Foundation (RSF) grant 17-15-01495

References:

1. J.Benichou et al. (2012). Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*.135(3):183-191.

2. M.Shugay et al. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol*. 11(11):e1004503.

3. D.V. Bagaev et al. (2016). VDJviz: a versatile browser for immunogenomics data. *BMC Genomics*. 17:453.

4. O.V. Britanova et al. (2016). Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *J Immunol*. 196(12):5005-5013.