Author/s:
Shugay, M;Bagaev, DV;Zvyagin, IV;Vroomans, RM;Crawford, JC;Dolton, G;Komech, EA;Sycheva, AL;Koneva, AE;Egorov, ES;Eliseev, AV;Van Dyk, E;Dash, P;Attaf, M;Rius, C;Ladell, K;McLaren, JE;Matthews, KK;Clemens, EB;Douek, DC;Luciani, F;Van Baarle, D;Kedzierska, K;Kesmir, C;Thomas, PG;Price, DA;Sewell, AK;Chudakov, DM

Title:
VDJdb: A curated database of T-cell receptor sequences with known antigen specificity

Date:
2018-01-01

Citation:
Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B. ,... Chudakov, D. M. (2018). VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Research, 46 (D1), pp.D419-D427. https://doi.org/10.1093/nar/gkx760.

Persistent Link:
https://hdl.handle.net/11343/222151

Published online 1 September 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D419–D427***
doi: 10.1093/nar/gkx760

VDJdb: a curated database of T-cell receptor sequences with known antigen specificity

Mikhail Shugay[1,2,3,4,5,†], Dmitriy V. Bagaev[3,†], Ivan V. Zvyagin[1,3], Renske M. Vroomans[6], Jeremy Chase Crawford[7], Garry Dolton[8], Ekaterina A. Komech[1,3], Anastasiya L. Sycheva[3], Anna E. Koneva[3], Evgeniy S. Egorov[1,3,5], Alexey V. Eliseev[1,3], Ewald Van Dyk[6], Pradyot Dash[7], Meriem Attaf[8], Cristina Rius[8], Kristin Ladell[8], James E. McLaren[8], Katherine K. Matthews[8], E. Bridie Clemens[9], Daniel C. Douek[10], Fabio Luciani[11], Debbie van Baarle[12], Katherine Kedzierska[9], Can Kesmir[6], Paul G. Thomas[7], David A. Price[8,10,13], Andrew K. Sewell[8,13] and Dmitriy M. Chudakov[1,2,3,4,5,*]

[1]Pirogov Russian National Research Medical University, Moscow 117997, Russia, [2]Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Institute of Science and Technology, Moscow 143028, Russia, [3]Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow 117997, Russia, [4]Central European Institute of Technology, Brno 60177, Czech Republic, [5]Nizhny Novgorod State Medical Academy, Nizhny Novgorod 603950, Russia, [6]Theoretical Biology and Bioinformatics, Science Faculty, Utrecht University, Utrecht 3512 JE, The Netherlands, [7]Department of Immunology, St. Jude's Children's Research Hospital, Memphis, TN 38105, USA, [8]Division of Infection and Immunity, Cardiff University School of Medicine, Cardiff CF14 4XN, UK, [9]Department of Microbiology and Immunology, University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Parkville VIC 3010, Australia, [10]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA, [11]Viral Immunology Systems Program, Kirby Institute, School of Medical Sciences, University of New South Wales, Kensington NSW 2052, Australia, [12]Center for Immunology of Infectious Diseases and Vaccines, National Institute for Public Health and the Environment, Bilthoven 3720 BA, The Netherlands and [13]Systems Immunity Research Institute, Cardiff University School of Medicine, Cardiff CF14 4XN, UK

Received June 23, 2017; Revised August 7, 2017; Editorial Decision August 16, 2017; Accepted August 17, 2017

ABSTRACT

The ability to decode antigen specificities encapsulated in the sequences of rearranged T-cell receptor (TCR) genes is critical for our understanding of the adaptive immune system and promises significant advances in the field of translational medicine. Recent developments in high-throughput sequencing methods (immune repertoire sequencing technology, or RepSeq) and single-cell RNA sequencing technology have allowed us to obtain huge numbers of TCR sequences from donor samples and link them to T-cell phenotypes. However, our ability to annotate these TCR sequences still lags behind, owing to the enormous diversity of the TCR repertoire and the scarcity of available data on T-cell specificities. In this paper, we present VDJdb, a database that stores and aggregates the results of published T-cell specificity assays and provides a universal platform that couples antigen specificities with TCR sequences. We demonstrate that VDJdb is a versatile instrument for the annotation of TCR repertoire data, enabling a concatenated view of antigen-specific TCR sequence motifs. VDJdb can be accessed at https://vdjdb.cdr3.net and https://github.com/antigenomics/vdjdb-db.

INTRODUCTION

VDJdb is a comprehensive database of antigen-specific T-cell receptor (TCR) sequences acquired by manual processing of published studies that report the ligand specificities of defined T-cell clonotypes (1). The primary goal of VDJdb is to facilitate access to existing information on TCR antigen specificities, i.e. the ability to recognize known epitopes presented by known major histocompatibility complex (MHC)

*To whom correspondence should be addressed. Tel: +7 499 742 81 22; Fax: +7 495 330 70 56; Email: chudakovdm@mail.ru
†These authors contributed equally to the paper as first authors.
Disclaimer: The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

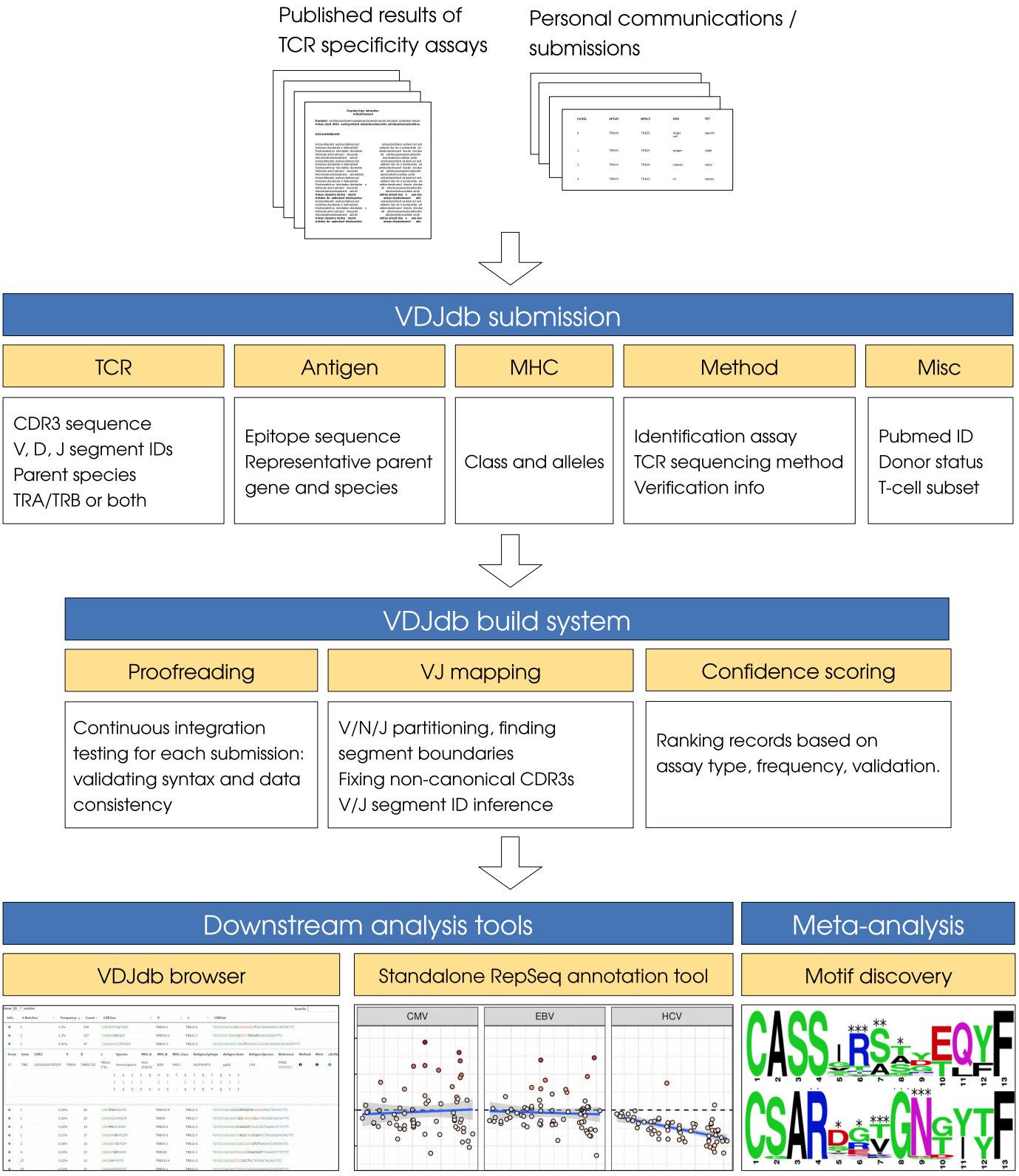

Figure 1. VDJdb overview. The VDJdb database aggregates published and communicated TCR sequences with known antigen specificities. Each VDJdb submission contains descriptions of the TCR α and/or β rearrangement (including the amino acid sequence of the somatically rearranged CDR3 loop), the cognate epitope (peptide sequence, representative parent gene and species) and the restricting MHC allotype, together with methodological details and other metadata. Submissions are checked for syntax errors and data consistency, V and J segments are mapped to the CDR3 sequences to define germline boundaries (V/J segments are inferred if not available in the submission), and a record confidence score is computed based on the methodological metadata. The database can be explored using the VDJdb browser web application, and RepSeq samples can be annotated using a standalone command-line tool. Meta-analysis of the VDJdb database can also facilitate the discovery of antigen-specific TCR motifs.

Table 1. VDJdb summary

Species	MHC	Gene	Records	Paired	Epitopes	Publications
Homo Sapiens	MHCI	TRA	481	310	65	55
H. Sapiens	MHCI	TRB	3001		99	75
H. Sapiens	MHCII	TRA	138	17	12	10
H. Sapiens	MHCII	TRB	451		14	11
Macaca Mulatta	MHCI	TRA	74	0	1	1
M. Mulatta	MHCI	TRB	1313		3	2
Mus Musculus	MHCI	TRA	8	7	8	8
M. Musculus	MHCI	TRB	8		8	8
M. Musculus	MHCII	TRA	8	8	8	6
M. Musculus	MHCII	TRB	9		8	6
Overall			5491	413	132	105

The number of TCR sequence records, epitopes and publications for each species, MHC class and TCR chain in the database. The 'Paired' column displays the number of records with known TCR α–β pairs. Statistics are calculated based on compiled VDJdb data from the 2 February 2017 database release.

Table 2. VDJdb scoring evaluation

VDJdb score	Multiple samples	Multiple studies	Count	Percent
0	–	–	1709/1918	89%
	+	–	166/1918	9%
	+	+	43/1918	2%
1	–	–	751/970	77%
	+	–	163/970	17%
	+	+	56/970	6%
2	–	–	211/348	61%
	+	–	27/348	8%
	+	+	110/348	32%
3	–	–	87/138	63%
	+	–	9/138	7%
	+	+	42/138	30%

The number and percentage of TCR:pMHC records that were reported only once, or found in multiple independent samples and/or studies for each VDJdb score value (from 0 to 3, higher score indicates higher confidence). Only human TCR β sequence records were considered.

class I and II molecules. Our mission is to aggregate TCR specificity information on a continuous basis and establish a curated repository to store these data in the public domain.

During the initial acquisition phase, we paid close attention to the associated metadata, which provide an indication of assay reliability and contribute to an in-depth understanding of TCR interactions with peptide–MHC (pMHC) complexes. An overview of the VDJdb database is provided in Figure 1, starting from data acquisition and proofreading routines to downstream analysis tools such as VDJdb browser and a standalone command-line utility that can be used to annotate large datasets containing TCRs with unknown specificities.

VDJdb complements existing popular immunogenetic resources by linking TCR sequences with their pMHC ligands. The IMGT database (http://www.imgt.org/, (2)) only stores germline TCR sequence information, while iEDB (http://www.iedb.org/, (3)) focuses on antigenic peptide epitopes without any linkage to cognate TCR sequences. VDJdb also complements another recently published TCR sequence annotation resource, the McPAS-TCR database (4). McPAS-TCR lists associations between TCR sequences and various pathologies, while VDJdb provides a more epitope-centric approach to TCR annotation, featuring a comprehensive description of TCR:pMHC interactions that largely disregards the underlying biological context.

Our rationale for developing VDJdb was closely allied to the advent of high-throughput sequencing of immune repertoires (RepSeq), a technique that allows the rapid acquisition of millions of distinct TCR sequences from individual samples (5). A lack of means to interrogate the antigen specificities of sequence-defined TCRs currently limits the information that can be extracted from these vast datasets. The need for VDJdb is highlighted by an emerging recognition that genetic associations with disease outcome can be refined by overlaying somatic rearrangements within the antigen-driven repertoire. For example, specific clonotypes have been shown to modulate the protective effects of human leukocyte antigen (HLA)-B*27 in HIV-1 infection (6) and Macaca mulatta (Mamu)-A*01 in SIV infection (7). The ability to uncover similar links across a range of immune-mediated conditions will be facilitated by a systematic approach that pairs TCR specificities with high-throughput screening technologies such as RepSeq (8).

RESULTS

Data sources and processing

TCR specificity assays. Published studies describing TCR specificity are the primary source of content for VDJdb. Most of the current records describe antigen-specific TCR sequences extracted from pMHC multimer-labeled T-cell populations (9). These assays are generally reliable if accompanied by high resolution sorting and good flow cytometry practice (10). Other records describe antigen-specific TCR sequences identified using functional readouts, including *in*

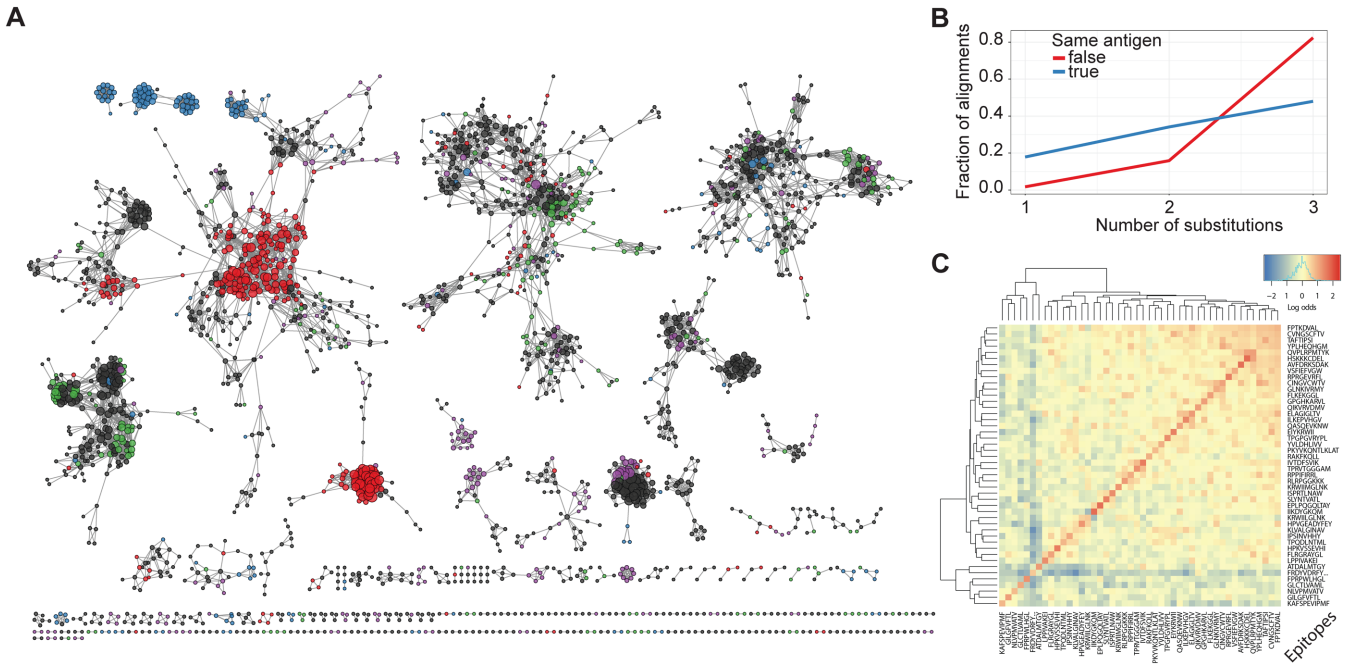

Figure 2. Similarity of TCR sequences specific for defined antigens. (**A**) The network of VDJdb records constructed using hamming distances computed for pairs of CDR3 amino acid sequences. Edges (alignments) connect sequences that differ by up to three amino acid substitutions. Nodes are colored by epitope: red, FRDYVDRFYKTLRAEQASQE (HIV-1/Gag); blue, GLCTLVAML (EBV/BMLF1); green, KRWIILGLNK (HIV-1/Gag); purple, NLVPMVATV (CMV/pp65); black, other epitopes. Node size is scaled by degree. Only human TCR β records were considered. Number of nodes, 2300; number of edges, 9651. (**B**) Frequency of alignments with a given number of substitutions for TCRs specific for the same (blue) or different epitopes (red). Number of same epitope alignments, 5285; number of different epitope alignments, 4366. (**c**) Heatmap showing the normalized number of alignments between each pair of epitope specificities. The diagonal indicates alignments within the same epitope specificity. Normalization was performed by dividing each entry of the alignment count matrix by the product of the corresponding row and column sums.

vitro assays based on T-cell recognition of targets expressing defined pMHC molecules (11). Additional verification steps, including post-sort analysis and the expansion and retesting of T-cell clones, are listed in the associated metadata. A summary of available VDJdb records is provided in Table 1.

Grooming reported TCR sequences. Common problems with published TCR specificity data include a lack of V and J segment specification, and incomplete (i.e. lacking conserved CDR3 residues) or excessive (e.g. including the conserved F residue in the J segment FGXG motif) CDR3 sequences. Such database records greatly complicate direct comparisons with immune repertoire sequencing data and the computation of summary statistics. We have therefore implemented an additional data processing step that involves: (i) removing excessive or adding missing germline CDR3 residues; (ii) resolving V and J segments by mapping to the germline database; and (iii) highlighting records that could not be repaired or do not match the V/J segment assignments reported by the authors.

Sequence assignment reliability

False-positive identification of antigen-specific TCR sequences can arise from several methodological issues, including non-specific pMHC multimer staining and low resolution cell sorting (9,10). We have therefore implemented a confidence scoring system that accounts for: (i) the frequency of the TCR clone or clonotype in the acquired sample of antigen-specific TCRs; (ii) the reliability of the TCR sequencing method (Sanger sequencing, high-throughput sequencing or single-cell sequencing); and (iii) the use of additional validation procedures (e.g. functional assays). The confidence scoring system implemented for VDJdb categorizes records into four different groups, from low confidence/no information (score = 0) to very high confidence (score = 3), and largely corresponds with the probability of a given TCR:pMHC pair being independently validated in different donors and/or studies (Table 2). The scoring is described in the database specification [https://github.com/antigenomics/vdjdb-db/blob/master/README.md].

Database infrastructure and analysis tools

Database specification, hosting and data submission. A detailed specification of the VDJdb database is available online [https://github.com/antigenomics/vdjdb-db/blob/master/README.md], including information related to the TCR sequence, the cognate epitope and the restricting MHC allotype, together with methodological outlines and associated metadata (Figure 1). An XLS template for database submission is also available in the VDJdb repository. VDJdb storage, submission and proofreading infrastructure is centered around the GitHub web-based repository hosting service (https://github.com). Processed and pending publications are listed in the issues section, database submissions are handled as pull requests,

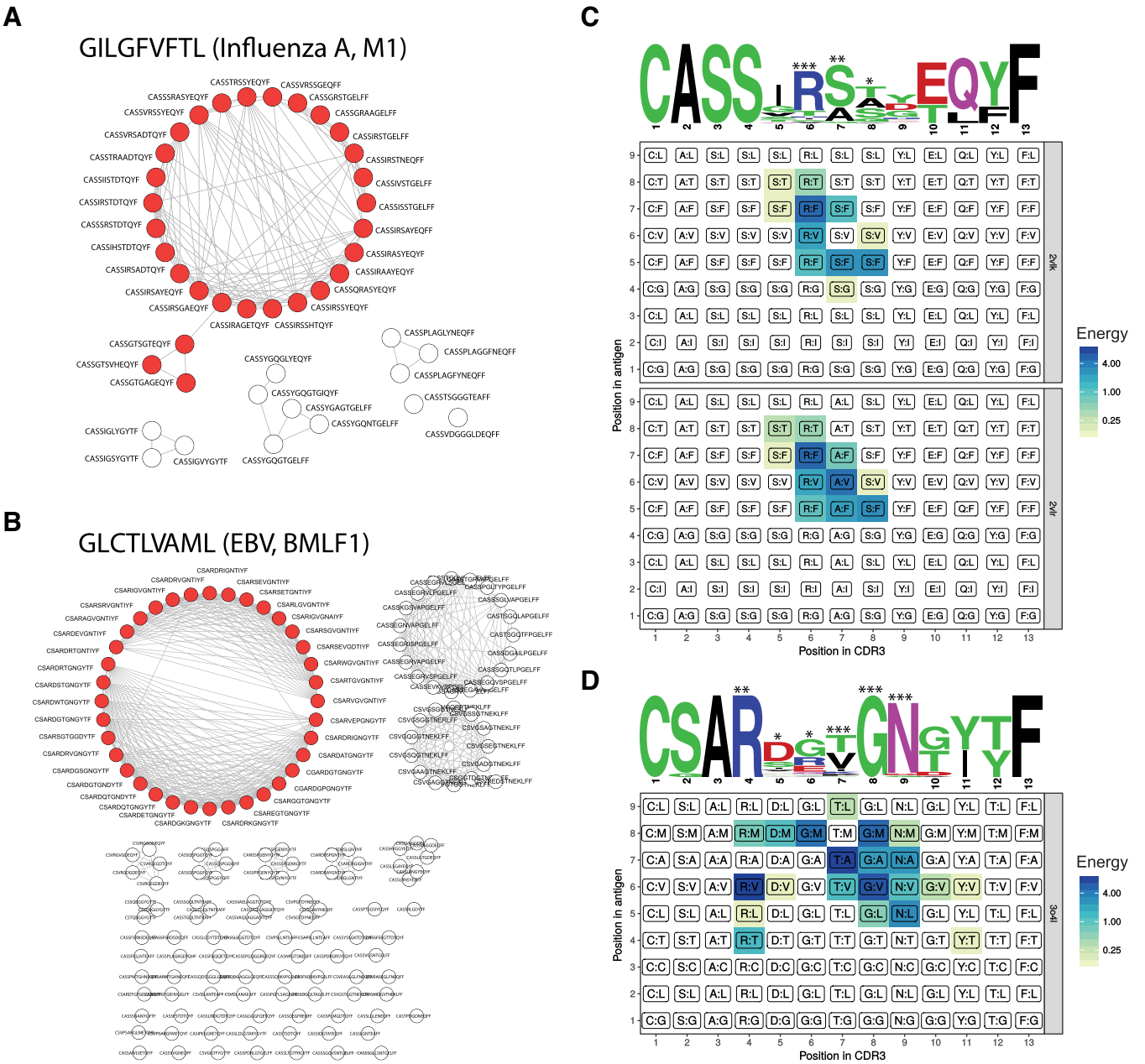

Figure 3. CDR3 motifs discovered from comparative analyses of VDJdb records. (**A** and **B**) Networks of pairwise alignments of GILGFVFTL-specific (A) and GLCTLVAML-specific (B) TCR β CDR3 sequences with up to three amino acid substitutions (no indels allowed). Nodes from the largest connected subnetworks (29 for GILGFVFTL and 36 for GLCTLVAML) used for motif discovery are shown in red. (**C** and **D**) TCR β CDR3 amino acid sequence logos and contact energy matrices obtained from available TCR:pMHC structural data. Sequence logos generated using WebLogo (http://weblogo.berkeley.edu/logo.cgi) show the relative frequency of each amino acid at each given position, and the height of each amino acid stack is scaled by the information content at each given position. Contact matrices are colored according to the interaction energies for each pair of CDR3 and peptide antigen residues (single-point energies were computed using GROMACS, value negated), and facet headers denote the Protein Data Bank IDs. Stars above the sequence logos show the number of accessible peptide antigen residues for each CDR3 residue, computed by counting peptide antigen residues closer than 5 Å to each given CDR3 residue.

and proofreading uses TravisCI continuous integration (https://travis-ci.org). Scripts required to assemble the database from the list of submissions, proofread, fix CDR3 sequences and assign confidence scores are included in the VDJdb repository.VDJdb web interface. A web-based GUI application for VDJdb browsing is available at https://vdjdb.cdr3.net/. The VDJdb browser allows the database to be queried for specific antigens or TCR sequences, and filtered based on parent species, V/J segments, MHC allotypes, antigen host species and/or TCR specificity assay-related metadata. Results can be exported as CSV and/or XLS tables.

VDJdb annotation tool. A standalone VDJdb annotation tool [https://github.com/antigenomics/vdjdb-standalone]

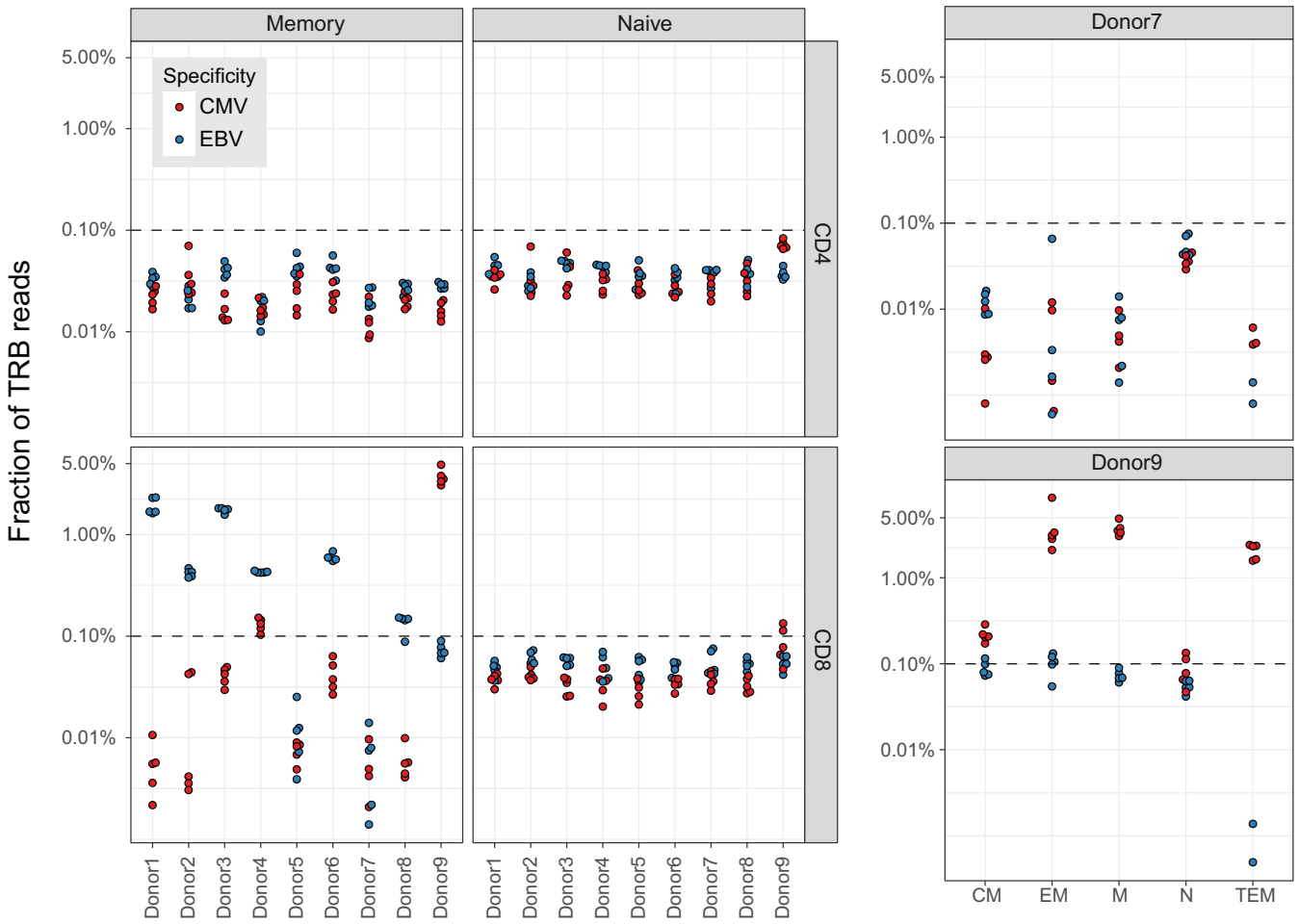

Figure 4. Annotation of TCR β (TRB) repertoires obtained from a study of naive (N) and memory (M) CD4 and CD8 T-cells (16). Five independent peripheral blood samples were analyzed per donor. The plot shows the fraction of TRB reads with matches to known CMV-specific or EBV-specific TCR β sequences in VDJdb. A detailed analysis of the memory CD8 T-cell subset in donor 7 (CMV-seronegative) and donor 8 (CMV-seropositive) is shown on the right. CM, central memory; EM, effector memory; TEM, terminally differentiated effector memory. The dashed line shows an ad hoc threshold of 0.1%, corresponding to the upper limit of specific TRB reads among naive T-cells.

was developed using the VDJtools framework (12). The tool can be used to annotate RepSeq samples processed through a variety of different software applications. It also generates comprehensive lists of clonotype matches, as well as annotation summary tables, and supports both exact CDR3 matching and hamming distance-based searches.

VDJdb application examples

Network structure of VDJdb records. Aggregated VDJdb records can be visualized as a network by constructing a graph with nodes corresponding to CDR3 amino acid sequences and edges connecting CDR3 amino acid sequences that differ by up to a fixed number of substitutions. The graph of currently available human TCR β CDR3 sequences contains 2300 nodes and 9651 edges, and suggests higher connectivity for records associated with the same antigen (Figure 2A, network layout and graphing performed using Cytoscape (13)). Of note, the hamming distance between CDR3 sequences that correspond to the same antigen is lower than the hamming distance between random CDR3 sequence pairs (Figure 2B), and TCRs specific for distinct antigens cluster neatly using the number of CDR3 pairs with hamming distances of <3 substitutions (Figure 2C).

TCR motif inference. The clonotypic architecture of antigen-specific T-cell populations is highly variable, ranging from epitopes that recruit a broad repertoire of diverse TCRs to epitopes that recruit a narrow repertoire of highly conserved TCRs, which can be exploited by viral escape mutations (14). Several epitopes in the VDJdb database are associated with a high number of unique TCR β CDR3 sequences, potentially enabling the identification of epitope-specific amino acid residues on the basis of CDR3 sequence similarity. Basic motif inference can be performed by selecting the most frequent connected subgraph for a given antigen and constructing length-identical position weight matrices (PWMs) (15). As an example, Figure 3A and B show CDR3 sequence graphs for the GILGFVFTL epitope from the influenza A virus matrix protein and the GLCTLVAML epitope from the Epstein–Barr virus BMLF1 protein. The selected sequences display a high degree of similarity and combine to generate a PWM with defined hot-spot posi-

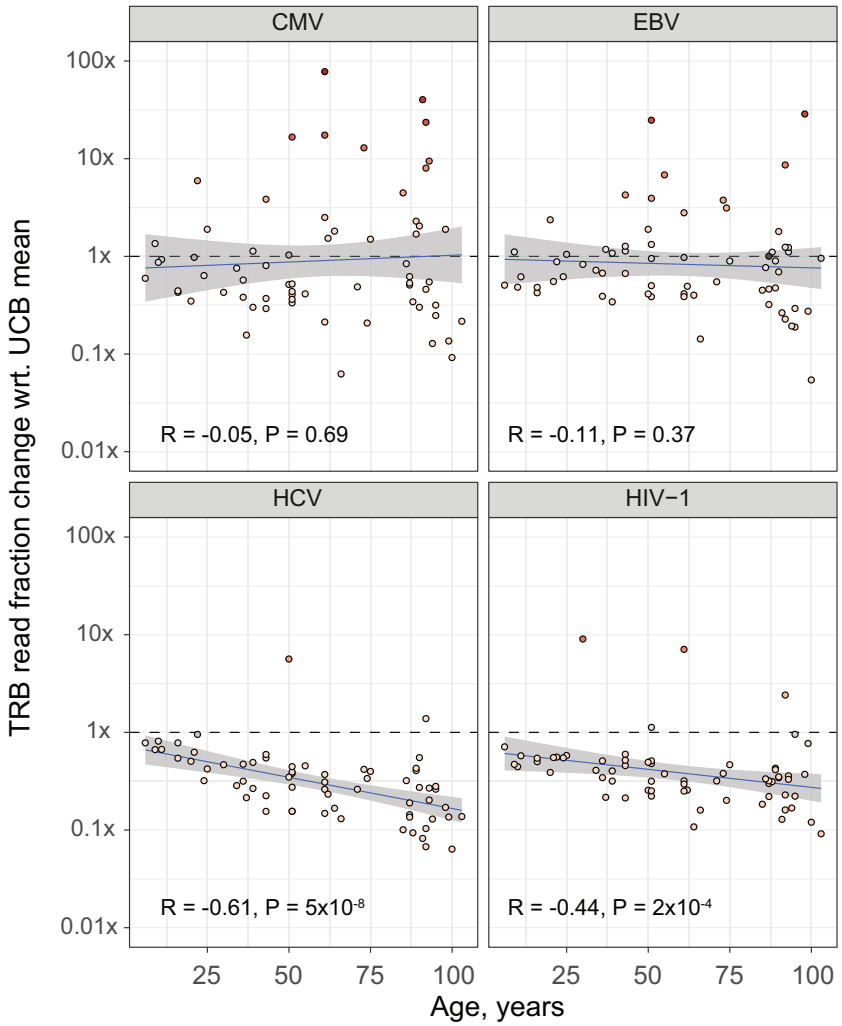

Figure 5. Abundance of TCR β (TRB) sequences specific for common (CMV and EBV) or less common persistent viruses (HCV and HIV) in peripheral blood samples from healthy donors of various ages ($n = 65$) (17). The plot shows the fraction of specific TRB reads divided by the mean value observed in umbilical cord blood (UCB) samples ($n = 8$). Z-scores were computed by comparing the TRB read fraction value in each donor with the corresponding mean and standard deviation values in UCB samples.

tions for each epitope specificity (Figure 3C and D). Of note, the CDR3 positions with the highest information content (neglecting germline bias from the V and J segments) in the PWMs juxtapose with TCR β CDR3:pMHC contacts defined by X-ray crystallography (Protein Data Bank accession codes: GILGFVFTL, 2vlk and 2vlr; GLCTLVAML, 3o4l).

Annotation of immune repertoire sequencing data. To demonstrate the utility of VDJdb as a tool for the annotation of large-scale RepSeq data, we selected two recently published datasets: (i) TCR sequences from CD4 and CD8 T-cell subsets isolated from donors stratified for age and CMV status (16); and (ii) pooled TCR sequences from an aging study (17).

In the first study, matching of CMV-specific and EBV-specific TCR β (TRB) sequences revealed a clear enrichment of specific variants in the memory CD8 T-cell subset, while the fraction of TRB reads that matched VDJdb records in the CD4 and naive CD8 T-cell subsets fell below an *ad hoc* threshold of 0.1% (Figure 4, left panel). A detailed subanalysis further showed that CMV-specific CD8 T-cells were present in different memory compartments in a CMV-seropositive donor but not in a CMV-seronegative donor (Figure 4, right panel).

In the second study, an analysis of specific TCR abundance in peripheral blood showed clear age-related trends for common and rare antigens. Across the cohort, a general decrease in the number of naive T-cells (17) was compensated by clonal expansions specific for common persistent viruses such as CMV and EBV (18), while the overall T-cell pool was depleted of clonotypes specific for rarer persistent viruses such as HCV and HIV (Figure 5). This result supports the notion that aging is associated with a progressive loss of clonal diversity that negatively impacts the ability to mount *de novo* immune responses (19).

DISCUSSION

Starting from the February 2017 release, VDJdb features more than 5000 records of mouse, monkey and human TCR sequences aggregated from more than 100 published studies, collectively spanning more than 100 different epitopes presented by MHC class I or class II. Various metadata are stored in addition to assess the quality of reported specificity assays and provide a confidence score for each database record. A proofreading step that automates TCR sequence annotation across the VDJ junction has also been included to ensure the consistency of records derived from studies using different nomenclatures and reporting styles.

As demonstrated above, VDJdb is a versatile resource that can be used to annotate large-scale TCR repertoire sequencing datasets and infer antigen-specific TCR motifs. Accordingly, VDJdb can be used as a benchmark for novel machine learning strategies designed to identify targeted antigens from primary TCR sequence data (20,21). Such approaches (e.g. performing fuzzy matching of TCR sequence motifs instead of relying on exact sequence matches) can extend coverage of the database and potentially enable comprehensive annotation of the entire set of possible TCR sequences, which can reach far beyond the estimate of 10^10 variants for human TCR β chains [https://arxiv.org/abs/1604.00487]. In addition, VDJdb can serve as one of the building blocks in a platform that integrates TCR specificity and peptide binding affinity data to solve one of the most ambitious and challenging tasks in the field, i.e. predicting T-cell recognition of a pMHC complex from the amino acid sequence of the TCR. The latter promises to increase the overall utility of immune repertoire sequencing technology in basic immunological research (22). Numerous applications can also be envisaged in translational settings. For example, accurate *in silico* predictions of TCR specificity would greatly facilitate cancer immunotherapy studies that rely on the adoptive transfer of tumor-specific T-cells (23).

The future development of VDJdb must interface with the emerging trend of switching from conventional (Sanger) sequencing to high-throughput techniques for the analysis of antigen-specific TCRs. Although sequencing the entire repertoire of an antigen-selected T-cell population will greatly increase method yield and allow the capture of low-frequency variants, it will eventually lead to the problem of distinguishing truly enriched TCR sequences from cellular contaminants and molecular artefacts. Model experiments incorporating rigorous controls, such as parallel comparisons of different sequencing methodologies with biological and procedural replicates, will therefore be required to assess reliability and develop guidelines for the management of datasets generated across diverse experimental and technical platforms.

DATA AVAILABILITY

VDJdb dataset is freely available under the Creative Commons Attribution-NoDerivatives 4.0 International license at https://github.com/antigenomics/vdjdb-db. VDJdb compilation and proofreading scripts are freely available under the Apache 2.0 license.

FUNDING

Pirogov Russian National Research Medical University; European Union's Horizon 2020 Research and Innovation Programme [№633592]; Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 [LQ1601]; Wellcome Trust Senior Investigator Awards (to D.A.P., A.K.S.); Cancer Research Wales PhD Studentship (to C.R.). Funding for open access charge: Skolkovo Institute of Science and Technology.
Conflict of interest statement. None declared.

REFERENCES

1. Bacher,P. and Scheffold,A. (2013) Flow-cytometric analysis of rare antigen-specific T cells. *Cytometry A*, **83**, 692–701.
2. Lefranc,M.P. (2003) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **31**, 307–310.
3. Vita,R., Overton,J.A., Greenbaum,J.A., Ponomarenko,J., Clark,J.D., Cantrell,J.R., Wheeler,D.K., Gabbard,J.L., Hix,D., Sette,A. *et al.* (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, **43**, D405–D412.
4. Tickotsky,N., Sagiv,T., Prilusky,J., Shifrut,E. and Friedman,N. (2017) McPAS-TCR: a manually-curated catalogue of pathology-associated T-cell receptor sequences. *Bioinformatics*, doi:10.1093/bioinformatics/btx286.
5. Benichou,J., Ben-Hamo,R., Louzoun,Y. and Efroni,S. (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, **135**, 183–191.
6. Ladell,K., Hashimoto,M., Iglesias,M.C., Wilmann,P.G., McLaren,J.E., Gras,S., Chikata,T., Kuse,N., Fastenackels,S., Gostick,E. *et al.* (2013) A molecular basis for the control of preimmune escape variants by HIV-specific CD8+ T cells. *Immunity*, **38**, 425–436.
7. Price,D.A., Asher,T.E., Wilson,N.A., Nason,M.C., Brenchley,J.M., Metzler,I.S., Venturi,V., Gostick,E., Chattopadhyay,P.K., Roederer,M. *et al.* (2009) Public clonotype usage identifies protective Gag-specific CD8+ T cell responses in SIV infection. *J. Exp. Med.*, **206**, 923–936.
8. Newell,E.W., Sigal,N., Nair,N., Kidd,B.A., Greenberg,H.B. and Davis,M.M. (2013) Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat. Biotechnol.*, **31**, 623–629.
9. Dolton,G., Tungatt,K., Lloyd,A., Bianchi,V., Theaker,S.M., Trimby,A., Holland,C.J., Donia,M., Godkin,A.J., Cole,D.K. *et al.* (2015) More tricks with tetramers: a practical guide to staining T cells with peptide-MHC multimers. *Immunology*, **146**, 11–22.
10. Chattopadhyay,P.K., Melenhorst,J.J., Ladell,K., Gostick,E., Scheinberg,P., Barrett,A.J., Wooldridge,L., Roederer,M., Sewell,A.K. and Price,D.A. (2008) Techniques to improve the direct ex vivo detection of low frequency antigen-specific CD8+ T cells with peptide-major histocompatibility complex class I tetramers. *Cytometry A*, **73**, 1001–1009.
11. Riddell,S.R. and Greenberg,P.D. (1990) The use of anti-CD3 and anti-CD28 monoclonal antibodies to clone and expand human antigen-specific T cells. *J. Immunol. Methods*, **128**, 189–201.
12. 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. *et al.* (2015) VDJtools: unifying post-analysis of T-cell receptor repertoires. *PLoS Comput. Biol.*, **11**, e1004503.
13. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
14. Price,D.A., West,S.M., Betts,M.R., Ruff,L.E., Brenchley,J.M., Ambrozak,D.R., Edghill-Smith,Y., Kuroda,M.J., Bogdan,D., Kunstman,K. *et al.* (2004) T-cell receptor recognition motifs govern immune escape patterns in acute SIV infection. *Immunity*, **21**, 793–803.
15. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

16. Qi,Q., Liu,Y., Cheng,Y., Glanville,J., Zhang,D., Lee,J.Y., Olshen,R.A., Weyand,C.M., Boyd,S.D. and Goronzy,J.J. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 13139–13144.
17. 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. *et al.* (2016) Dynamics of individual T-cell repertoires: from cord blood to centenarians. *J. Immunol.*, **196**, 5005–5013.
18. Wallace,D.L., Masters,J.E., De Lara,C.M., Henson,S.M., Worth,A., Zhang,Y., Kumar,S.R., Beverley,P.C., Akbar,A.N. and Macallan,D.C. (2011) Human cytomegalovirus-specific CD8(+) T-cell expansions contain long-lived cells that retain functional capacity in both young and elderly subjects. *Immunology*, **132**, 27–38.
19. De Martinis,M., Franceschi,C., Monti,D. and Ginaldi,L. (2005) Inflamm-ageing and lifelong antigenic load as major determinants of ageing rate and longevity. *FEBS Lett.*, **579**, 2035–2039.
20. Dash,P., Fiore-Gartland,A.J., Hertz,T., Wang,G.C., Sharma,S., Souquette,A., Crawford,J.C., Clemens,E.B., Nguyen,T.H.O., Kedzierska,K. *et al.* (2017) Quantifiable predictive features define epitope-specific T-cell receptor repertoires. *Nature*, **547**, 89–93.
21. Glanville,J., Huang,H., Nau,A., Hatton,O., Wagar,L.E., Rubelt,F., Ji,X., Han,A., Krams,S.M., Pettus,C. *et al.* (2017) Identifying specificity groups in the T-cell receptor repertoire. *Nature*, **547**, 94–98.
22. Heather,J.M., Ismail,M., Oakes,T. and Chain,B. (2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief. Bioinform.*, doi:10.1093/bib/bbw138.
23. Thomas,S., Stauss,H.J. and Morris,E.C. (2010) Molecular immunology lessons from therapeutic T-cell receptor gene transfer. *Immunology*, **129**, 170–177.