Bioinformatics, 33(18), 2017, 2924–2929
doi: 10.1093/bioinformatics/btx286
Advance Access Publication Date: 8 May 2017
Original Paper

OXFORD

Databases and ontologies

McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences

Nili Tickotsky[1], Tal Sagiv[1], Jaime Prilusky[2], Eric Shifrut[1] and Nir Friedman[1,*]

[1]Department of Immunology and [2]Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot 76100, Israel

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on February 2, 2017; revised on April 14, 2017; editorial decision on April 21, 2017; accepted on May 2, 2017

Abstract

Motivation: While growing numbers of T cell receptor (TCR) repertoires are being mapped by high-throughput sequencing, existing methods do not allow for computationally connecting a given TCR sequence to its target antigen, or relating it to a specific pathology. As an alternative, a manually-curated database can relate TCR sequences with their cognate antigens and associated pathologies based on published experimental data.

Results: We present McPAS-TCR, a manually curated database of TCR sequences associated with various pathologies and antigens based on published literature. Our database currently contains more than 5000 sequences of TCRs associated with various pathologic conditions (including pathogen infections, cancer and autoimmunity) and their respective antigens in humans and in mice. A web-based tool allows for searching the database based on different criteria, and for finding annotated sequences from the database in users' data. The McPAS-TCR website assembles information from a large number of studies that is very hard to dissect otherwise. Initial analyses of the data provide interesting insights on pathology-associated TCR sequences.

Availability and implementation: Free access at http://friedmanlab.weizmann.ac.il/McPAS-TCR/.

Contact: nir.friedman@weizmann.ac.il

1 Introduction

T cells recognize antigens of foreign and self-origin using the T cell receptor (TCR), which is comprised of two chains (TCRα, TCRβ) that are generated through a random process of DNA rearrangement (Bassing *et al.*, 2002; Davis and Bjorkman, 1988). This random process can generate a huge and diverse repertoire of TCR sequences, which ensures functional adaptive immunity (Laydon *et al.*, 2015; Robins *et al.*, 2010). Responding T cell populations typically show profiles of preferred usage of a few TCR sequences (oligoclonal responses) (Kedzierska *et al.*, 2008). Moreover, specific pathology-related TCR sequences were detected across different individuals with the same pathology (Madi *et al.*, 2014; Mamedov *et al.*, 2011; Turner *et al.*, 2006; Venturi *et al.*, 2008). These findings indicate that TCR bias exists in pathological situations (Miles *et al.*, 2010; Serana *et al.*, 2009; Turner *et al.*, 2006; Venturi *et al.*, 2008), and the significance of this concept can be studied through the association of experimentally observed expanded TCR sequences with their respective pathologic conditions and antigens.

Although current progress allows for mapping of TCR repertoires using high-throughput sequencing (Calis and Rosenberg, 2014; Friedensohn *et al.*, 2017; Heather *et al.*, 2017; Mamedov *et al.*, 2013), existing methods do not allow for computationally connecting a given TCR sequence to the antigen that it binds and recognizes. Studies of disease-specific TCR sequences frequently try to get a broader perspective on their findings by comparing them with previous disease-related sequence data (Emerson *et al.*, 2017; Hughes *et al.*, 2003; Madi *et al.*, 2014; Mamedov *et al.*, 2011). However, such data is sparse and difficult to retrieve and assess. To overcome this limitation, we present McPAS-TCR, a manually curated database of TCR sequences that are associated with various pathologies and antigens, based on published literature.

We believe that McPAS-TCR fills an important unmet need, as it collects in one accessible website information from a large number of studies that is very hard to dissect otherwise. The database includes detailed information on each TCR sequence, such as details on the study in which it was found, the type of pathology with which it was associated, the related antigen protein/epitope (if known), and the way by which antigen specificity was determined. It provides valuable information that can support hypotheses and new studies in many areas of immunology, including vaccine design, autoimmunity and cancer immunotherapy.

2 Materials and methods

2.1 Data collection

To build a database of annotated TCRα and TCRβ sequences that were found in T cell clones associated with various diseases/conditions in humans and in mice, a PubMed query was conducted with variations of the following search terms: CDR3 sequence, sequencing, T cell receptor, T cell repertoire, T cell AND [Pathology X]. Further publications were added based on citations from papers that were found by the PubMed search. Only data from published work was included.

2.2 Data organization

Each entry in the database describes one TCR sequence identified in one study. TCR sequences were manually curated and characterized according to the following fields:

(1) TCR alpha and beta chains: The junctional region of the TCR, also known as the complementarity-determining region 3 (CDR3), is highly diverse and constitutes an important determinant of antigen recognition by T cells. We list for each TCR sequence its CDR3 region (amino-acid sequence), starting with the conserved cysteine at the 5’ end of the V segment, and ending with the conserved phenylalanine at the 3’ end of the J segment. V segment and J segment are presented if given in the original publication. Nomenclature follows the international ImMunoGeneTics information system (http://www.imgt.org) (Lefranc *et al.*, 1999). When a study used a different gene ontology system, it was converted to the IMGT nomenclature.

The database contains both TCRα and TCRβ sequences. Some entries contain both sequences, if this information was available. Nucleotide sequences are also presented if available.

As a rule, sequences are presented as given in the original paper. However, when the conserved cysteine amino acid at the beginning of the CDR3 sequence was missing in the original paper, it was added to the database description. The same applies to a missing conserved phenylalanine at the end of the CDR3 sequence. In addition, when the TRBJ gene was not mentioned in the article, it was reconstructed according to the CDR3 sequence, if possible, based on the IMGT terminology (Lefranc *et al.*, 1999). Such corrections are stated in the field ‘reconstructed J annotation’.

(2) Category: Sequences were classified to one of the following categories, based on the pathology with which they were associated: (A) Pathogens–including bacteria, viruses and parasites, (B) Autoimmune–including sequences identified in tissues/Tcells from human and mice with an autoimmune condition, (C) Cancer–including sequences identified in malignant tissues/Tcells of human origin, or in mice models of malignancies, (D) Allergy–including sequences identified in allergic reactions to various allergens and (E) Other–including sequences not classified to any of the above categories.

(3) Pathology: This field describes a disease or some sort of medical condition (e.g. EBV viral infection, rheumatoid arthritis, melanoma). This information enables the user to find all curated sequences related to a specific disease/pathological condition.

(4) Additional study details: This field adds additional relevant details pertaining to the study in which the sequence was identified. For example, EBV related sequences that were identified in patients following bone-marrow transplantation.

(5) Antigen identification method: In cases where the cognate antigen was identified, this field presents the method by which the association of the TCR with the antigen was detected. Selection of antigen-specific T cells using peptide-MHC multimers (Altman *et al.*, 1996) is labeled ‘1’. Selection of T cells that were activated by an antigen in culture is labeled ‘2’, with subcategories 2.1–2.5 corresponding to the type of antigen used for stimulation (2.1: stimulation with a peptide, 2.2: a whole protein, 2.3: a whole pathogen, 2.4: tumor cells, 2.5: other types of *in vitro* stimulation). Sequences revealed by directly sequencing T cells *ex-vivo* are labeled ‘3’. These include either Sanger sequencing of selected clones or the 50 most abundant sequences from a high-throughput sequencing experiment. These can be, for example, tumor infiltrating T cells, or T cells from synovial fluid of rheumatoid arthritis patients. These sequences are associated with a specific pathology, and are expected to be enriched with TCRs specific for relevant antigens, e.g. cancer neo-antigens, the identity of which is not known. This score enables users to filter the data according to their need and research question. (See further discussion in Section 3.1).

(6) NGS: This field specifies if the TCR sequence was identified using Next Generation Sequencing (NGS). It states either ‘Yes’, for use of NGS, or ‘No’ otherwise.

(7) Antigen protein: This field describes the antigen protein that the TCR targets, if described in the curated article. The Protein ID is the protein’s entry number in the UniProt Knowledgebase (UniProtKB, http://www.uniprot.org/uniprot/) (UniProt, 2015).

(8) Epitope peptide: This field describes the specific peptide epitope that is recognized by the TCR, if described in the curated article. The Epitope ID is the epitope’s entry in the Immune Epitope Database (IEDB, www.iedb.org) (Vita *et al.*, 2015).

(9) MHC: The major histocompatibility complex associated with the epitope, as described in the original article. For human TCRs, the HLA allele is presented, as described at the HLA Nomenclature website (http://hla.alleles.org/nomenclature/naming.html). For mouse TCRs, the MHC haplotype of the corresponding mouse strain is given.

(10) Tissue: This field describes the tissue from which the T cells were extracted.

(11) T cell type: Two fields describe the type of T cells from which the sequence was obtained: CD4/CD8, and additional details on the T cell type, e.g. T-regulatory, T-effector, or TIL (Tumor infiltrating lymphocyte).

(12) PubMed.ID: The PubMed Identifier (PMID) of the study’s entry in PubMed database of citations for biomedical literature at http://www.ncbi.nlm.nih.gov/pubmed . A link is provided so that the paper’s abstract is easily accessible.

2.3 Web interface

The web interface for McPAS-TCR was developed in R using the Shiny package from R studio (http://shiny.rstudio.com/). A remote web server (http://friedmanlab.weizmann.ac.il/McPAS-TCR/) is backing the R web interface. The graphical user interface (GUI) relies on the shinydashboard R package (https://rstudio.github.io/shinydashboard/index.html). Session-based user searches are included in the framework design. Multiple sequence searches are enabled by the shinyAce R package. The user can search not only for sequences that are identical to the query sequence, but also for similar sequences with varying edit distances. This feature is made possible through the stringdist R package (https://cran.r-project.org/web/packages/stringdist/index.html). The Levenshtein distance (Levenshtein 1966) is used to estimate similarity between the query and the annotated sequences. The search results can be filtered by specific values in the different fields, using the DT R package filtering feature (https://rstudio.github.io/DT/).

The integrated data can be downloaded in comma-separated values (CSV) format. Each entry in the database is linked to other public databases. Links include: publications from the National Center for Biotechnology Information (NCBI) PubMed site, pathologies as defined in the vocabulary of biomedical terms of the Medical Subject Headings (MESH) database of the U.S.A national library of medicine (NLM), antigen proteins are linked to the Uniprot protein database (UniProt, 2015) by their protein ID, and epitope peptides are linked to the IEDB database of T cell epitopes (Vita *et al.*, 2015).

3 Results

3.1 McPAS-TCR overview

The database currently contains ~5100 entries, curated from 118 publications, of TCR sequences that were found in T cells associated with various pathologic conditions in humans and in mice. ~75% of the sequences are from human data, and the rest come from mouse data. Each sequence is characterized by its V and J gene segments and by the amino acid sequence of the complementarity-determining region 3 (CDR3). In most cases only the TCRβ sequence is available, some entries contain the TCRα sequence, and some contain both. Other characteristics are included if found in the original paper, as described in the methods section.

We classified the pathologies into five categories: pathogens, autoimmune, cancer, allergy and other. Table 1 shows the number of entries in each category in our database. Each category contains different pathologies. Examples include T cells associated with pathogens, such as viral infections: Cytomegalovirus (CMV, $N=459$ entries), Epstein Barr virus (EBV, $N=474$) or influenza ($N=498$); T cells associated with autoimmune diseases (e.g. multiple sclerosis, $N=118$, type 1 diabetes, $N=532$); and T cells associated with malignancies (e.g. melanoma, $N=542$). We also include other relevant information such as the tissue of origin of the T cells, their type (CD4/CD8) and other characteristics (e.g. memory/regulatory, TIL, etc).

Relating an antigen to specific TCRs that recognize it is a main aim of this database. Different studies provide varying level of information regarding the antigens. This information is reflected by a number of fields in the database. The 'antigen identification method' describes the method by which the association between the antigen and TCR was determined. A value of '1' indicates detection by peptide-MHC multimers (Altman *et al.*, 1996), which is considered a gold-standard for detection of antigen-specific T cells. A value of '2' indicates isolation of T cells that respond to a specific antigen from *in vitro* cultures. Various types of antigens are used in these studies: a peptide antigen, an antigenic protein, a whole pathogen or cells from a tumor. This is indicated by sub-divisions 2.1–2.5, as described in the methods section. We also included in the database sequences associated with a pathology for which the antigen is not identified in the original study. These include, for example, tumor infiltrating T cells, T cells isolated from pancreases of type 1 diabetes patients, or T cells isolated from synovial fluid of rheumatoid arthritis patients. These sequences are labeled by '3' in our antigen identification method scheme. This information can enable users to filter the results based on the requested level of accuracy, while supplying information on TCR sequences that might be relevant in some pathologic scenarios, even if currently the specific antigen is not known. It should be noted, however, that in these cases (category '3') the sequences may come from bystander cells, not necessarily related to the pathology in a direct way.

Table 1. Number of TCRα and TCRβ sequences in McPAS-TCR per functional category, in mouse and human data

Category	Pathogens	Autoimmune	Cancer	Allergy	Total
Human TCRα	252	91	41	2	386[a]
Human TCRβ	1,897	886	804	259	3887
Mouse TCRα	105	52	68	0	254[b]
Mouse TCRβ	580	279	220	0	1194

[a] 270 human TCRα and TCRβ chains are paired.

[b] 148 mouse TCRα and TCRβ chains are paired.

The method by which the TCR sequences were obtained is indicated by the 'NGS' filed. In studies that use next generation sequencing (NGS), we included only the 50 most abundant TCR sequences from each study, assuming that these are the most relevant. These cases are indicated by the entry 'yes' in the NGS field. If sequences were identified by sequencing isolated T cell clones, they are marked by the entry 'no' in that filed. Information on MHC/HLA allele is also included in the database if available.

3.2 McPAS-TCR search options

The database can be searched using a web-based friendly user interface that provides various data analysis options. Users can search the database for specific sequences, and filter data according to specific values of the different fields. We also allow for search within long lists of sequences (as obtained from high-throughput TCR sequencing studies, for example), by uploading user text files and querying them against the database. Search results can be saved to a file for further analysis.

The search can return sequences from the database that are either identical to the user query, or those that are similar to the query up to a user-defined threshold (measured by the Levenshtein distance (Levenshtein 1966)–the minimal number of insertions/deletions/substitutions required to change one sequence to the other). Thus, users can find not only identical sequences, but also sequences that are similar to a given query sequence, which might indicate related antigen specificity.

Some examples for searching the database include:

(1) Find TCR sequences associated with a specific pathology/disease.

(2) Find TCR sequences associated with a specific antigen or epitope.

(3) Compare human CDR3 sequences with mouse CDR3 sequences associated with a specific pathology/disease.

(4) Find sequences from the database in lists of CDR3 sequences provided by the user (e.g. from high-throughput TCR sequencing data).

3.3 Update of McPAS-TCR

We intend to continue collecting more sequences to the database, and maintain it functional and updated. We currently include in the database TCRα and TCRβ sequences. TCRγ and TCRδ sequences can be added in the future, when such pathology related sequences will become available in the literature.

3.4 Applications

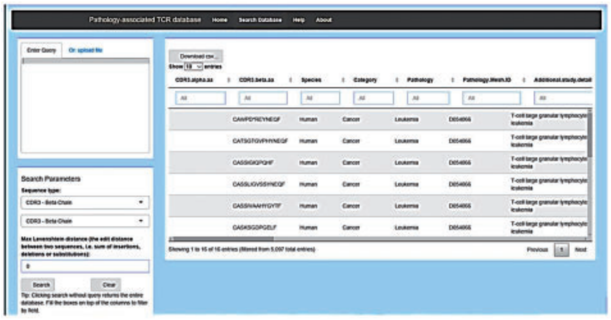
3.4.1 TCRs associated with leukemia

We searched the database for sequences that are related to leukemia. In the 'Search database' window, we typed 'leukemia' inside the query field above the 'Pathology' column. This gives three terms in the pop-up list: 'Leukemia', 'Acute myeloid leukemia' and 'Murine leukemia virus (MuLV)', which were all checked. The resulting search screen is shown in Figure 1. This search returned sequences from three studies: a study on T-large granular lymphocyte (T-LGL) leukemia (Clemente *et al.*, 2013), another study that sequenced TCRs from human CD8+ T cells specific for a leukemia derived antigenic peptide epitope (Hunsucker *et al.*, 2015) and one mouse study (Iwashiro *et al.*, 1993), in which T cell clones were stimulated by peptides from murine leukemia virus envelope proteins.

The first study used high-throughput TCR sequencing of T cells obtained from peripheral blood of T-LGL leukemia patients. Fourteen immunodominant clones from 11 patients are described in the paper and listed in the database, providing information on the sequence attributes of the clones (V, D, J segments, CDR3βaa and CDR3βnt sequences). We further searched our database for these 14 sequences, and found that one of them (CASSLIGVSSYNEQF) was previously identified in a study of CMV specific CD8 T cells (Miconnet *et al.*, 2011). In the T-LGL leukemia study, this sequence came from a dominant expanded clone that accounted for 10.9% of all TCRβ sequences in one of the patients. This example shows how the database can be used to link findings of different papers and foster new research hypotheses.

3.4.2 TCRs associated with a specific peptide epitope

McPAS-TCR allows for studying associations between specific epitopes and the TCRs that recognize them. As an example, we searched for CDR3 sequences associated with the Cytomegalovirus (CMV) immunodominant epitope NLVPMVATV, derived from the pp65 viral protein. In the 'Search Database' window we typed 'NLVPMVATV' in the sequence query field, and selected 'Epitope' from the available search parameters. The search returned 237 entries of TCRβ CDR3 sequences associated with that epitope, from 10 publications. For each sequence, the search results also provided the related parameters (Pubmed ID, TRBV and TRBJ, cell type etc.). Interestingly, we find in this list some TCRβ sequences that were identified in more than one study. One example is 'CASSLAPGATNEKLF', that has been found in six studies (Table 2). This CDR3β amino acid sequence was associated with one TRBJ gene segment (TRBJ1-4), but with three different TRBV segments in the different studies (TRBV7-6, TRBV5-1 and TRBV6-3). In one study, it was encoded by two different CDR3nt sequences. These findings indicate dominance of this public TCR, and suggest convergent recombination of sequences that were generated by different DNA recombination events and were selected based on their common antigen specificity.


Pathology-associated TCR database
Home
Search Database
Help
About
Enter Query
Or upload file
Search Parameters
Sequence type:
CDR3 - Beta Chain
CDR3 - Beta Chain
Max Levenshtein distance (the edit distance between two sequences, i.e. sum of insertions, deletions or substitutions):
0
Search
Clear
Tip: Clicking search without query returns the entire database. Fill the boxes on top of the columns to filter by field.
Download csv...
Show 10 entries

CDR3.alpha.aa	CDR3.beta.aa	Species	Category	Pathology	Pathology.Mesh.ID	Additional.study.detail
All	All	All	All	All	All	All
	CAWPDYREYNEQF	Human	Cancer	Leukemia	D054066	T-cell large granular lymphocyte leukemia
	CATSGTGVPHYNEQF	Human	Cancer	Leukemia	D054066	T-cell large granular lymphocyte leukemia
	CASSGGQPQHF	Human	Cancer	Leukemia	D054066	T-cell large granular lymphocyte leukemia
	CASSLIGVSSYNEQF	Human	Cancer	Leukemia	D054066	T-cell large granular lymphocyte leukemia
	CASSWAAHYGYTF	Human	Cancer	Leukemia	D054066	T-cell large granular lymphocyte leukemia
	CASASGDPGELF	Human	Cancer	Leukemia	D054066	T-cell large granular lymphocyte leukemia

Showing 1 to 15 of 15 entries (filtered from 5,097 total entries)

Fig. 1. The 'Search database' screen of the web-based tool. In this example, the term 'Leukemia' was chosen from a pop-up list in the 'Pathology' field. Shown are 6 out of 18 entries that were found. The user can view more entries by scrolling up-down, and further fields by scrolling left-right

3.4.3 TCRs identified in a number of studies

Next, we looked for TCRβ CDR3 sequences in the database that were identified in three or more studies. Twenty such cases were found, representing public sequences that are shared by many individuals with the same pathology (Madi *et al.*, 2014; Menezes *et al.*, 2007). One example is the TCRβ CDR3 sequence 'CSARDRTGNGYTF', which is associated with EBV (protein: EBV-BMLF-1, epitope: GLCTLVAML, HLA: HLA-A*01), and has been found in seven studies. Interestingly, there are ten sequences in the database that are at a Levenshtein distance of 1 from this EBV associated sequence (Table 3). These sequences were found in ten different studies, many of which more than once, and all were found to recognize the same EBV epitope. The ability to connect data from many studies and find highly similar but not identical sequences that share antigen specificity, can contribute for investigation of determinants of TCR specificity.

3.4.4 Search for annotated TCRs in high-throughput TCR-seq data

In a previous study, we have analyzed CDR3β sequences in TCR repertoires from spleens of young and old healthy C57BL/6 mice (Shifrut *et al.*, 2013). We now used McPAS-TCR to look for annotated CDR3β sequences in this data. To perform the search, we used the 'Upload file' option of the 'Search database' window to load lists

Table 2. Results obtained by searching the database for the CDR3β sequence 'CASSLAPGATNEKLF' that is associated with a CMV peptide epitope

Antigen identification method	MHC	TRBV	Pub-Med.ID
1[a]	HLA-A*02	TRBV7-6	21374820
1	HLA-A*02:01	TRBV7-6	19017975
1	HLA-A*02:01	TRBV5-1	21555537
1	HLA-A*02	TRBV7-6	16287711
1	HLA-A*02	TRBV6-3	16237109
1	HLA-A*02		26509579

[a] Binding to pMHC multimers.

Table 3. Results obtained by searching the database for sequences at a Levenshtein distance of 1 from the EBV-associated CDR3β sequence 'CSARDRTGNGYTF' (top row)

CDR3β sequence	Number of publications
CSARDRTGNGYTF	7
C_ARDRTGNGYTF	1
CSARDATGNGYTF	1
CSARDGTGNGYTF	4
CSARDKTGNGYTF	1
CSARDQTGNGYTF	3
CSARDSTGNGYTF	3
CSARDWTGNGYTF	1
CSARDRIGNGYTF	3
CSARDRKGNGYTF	1
CSARDRVGNGYTF	1

Note: All sequences were associated with the same EBV epitope. Modified positions and gaps are underlined.

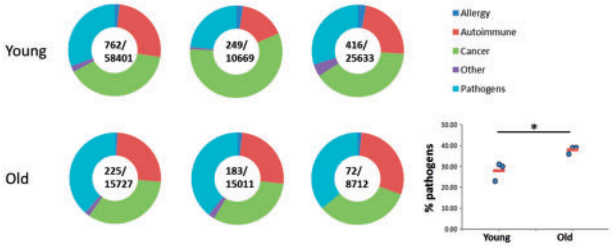


Fig. 2. Analysis of annotated sequences found in three young and three old healthy mice. Left: Each pie chart represents annotated sequences found in the repertoire of one mouse, and the sections represent the fraction of annotated sequences in each of the five functional categories. Numbers in the center indicate the number of annotated sequences found in each repertoire out of the total number of unique TCR sequences. Right: The fraction of pathogens-associated annotated sequences is higher in old mice relative to young ones (*P*-value = 0.02, two tailed *t*-test)

of the TCRβ CDR3 sequences found in the six mice from the previous study. We specified 'CDR3–Beta chain' in the 'Search parameters–sequence type' pop-up list and selected a Levenshtein distance of zero, i.e. exact match. The results are summarized in Figure 2, which shows the number of annotated sequences found in each mouse repertoire, and the fraction of sequences in each of the 5 functional categories. In this example, 1–2% of the sequences in these repertoires could be annotated using our database. We found that the fraction of pathogen-associated annotated sequences was significantly higher in old mice relative to young ones.

4 Discussion

McPAS-TCR is a comprehensive database of pathology-related TCR sequences that aims to fill an important unmet need of the immunology community by assembling existing information on pathology-related TCR sequences. This data is vital for deciphering the association between a given TCR sequence and the pathology with which it is associated, and in many cases also with the antigen that it binds and recognizes.

We demonstrated how this tool can be used to identify epitope specific TCR sequences across a large number of studies, and also to find similarities between sequences recognizing the same epitope. These large lists of epitope specific TCR sequences, obtained by integrating data from many resources, can be used to develop a better understanding of the relations between TCR sequences that recognize the same target.

The ability to search high-throughput TCR-seq data for annotated sequences can allow researchers for finding disease associated TCR sequences in their data, providing links to previous studies and evoking new hypotheses regarding sequences that co-occur in different settings. A notable example is the analysis of TCR repertoires of tumor infiltrating T cells (Cohen *et al.*, 2015; Shilyansky *et al.*, 1994), which can benefit from identification of TCRs with known targets.

We are aware of another resource (which has not been described in the literature) that contains TCR sequences collected from published manuscripts, and is searchable by users through a web-based interface (VDJdb, https://vdjdb.cdr3.net/). There is only a partial overlap in sequences and cited papers between McPAS-TCR and VDJdb, attesting to the very large diversity of antigen specific TCR repertoires, and the broad literature describing them. As both datasets are based on manual curation, they are complimentary and should allow users to cover the ever growing literature in this field. A major difference between the two resources is that VDJdb focuses on relating TCR sequences to their cognate epitopes, while McPAS-TCR focuses on the pathology with which a TCR sequence is associated. We believe that our approach is advantageous for appreciating the clinical relevance of TCR repertoires and will find applications in pre-clinical and clinical studies that study T cell responses. Of note, we include sequences that were found in defined pathologies even if their antigen is not known. In particular, TCRs associated with autoimmune diseases, or with tumors (TILs) are included in McPAS-TCR. Thus, McPAS-TCR can allow researchers to identify autoimmune-associated TCRs or TILs that are shared by individuals across different studies. Such links can in turn provide further information, for example correlation of TCRs with expression of specific neo-antigens, which can lead to new testable hypotheses.

Funding

This work was supported by a grant from the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation.

Conflict of Interest: none declared.

References

Altman,J.D. *et al.* (1996) Phenotypic analysis of antigen-specific T lymphocytes. *Science*, **274**, 94–96.

Bassing,C.H. *et al.* (2002) The mechanism and regulation of chromosomal V(D)J recombination. *Cell*, **109**(Suppl), S45–S55.

Calis,J.J. and Rosenberg,B.R. (2014) Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.*, **35**, 581–590.

Clemente,M.J. *et al.* (2013) Deep sequencing of the T-cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes. *Blood*, **122**, 4077–4085.

Cohen,C.J. *et al.* (2015) Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. *J. Clin. Invest.*, **125**, 3981–3991.

Davis,M.M. and Bjorkman,P.J. (1988) T-cell antigen receptor genes and T-cell recognition. *Nature*, **334**, 395–402.

Emerson,R.O. *et al.* (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet.*, **49**, 659–665.

Friedensohn,S. *et al.* (2017) Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends Biotechnol.*, **35**, 203–214.

Gao,F. and Wang,K. (2015) Ligation-anchored PCR unveils immune repertoire of TCR-beta from whole blood. *BMC Biotechnol.*, **15**, 39.

Heather,J.M. *et al.* (2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief Bioinform.*, DOI: 10.1093/bib/bbw138.

Hughes,M.M. *et al.* (2003) T cell receptor CDR3 loop length repertoire is determined primarily by features of the V(D)J recombination reaction. *Eur. J. Immunol.*, **33**, 1568–1575.

Hunsucker,S.A. *et al.* (2015) Peptide/MHC tetramer-based sorting of CD8(+) T cells to a leukemia antigen yields clonotypes drawn nonspecifically from an underlying restricted repertoire. *Cancer Immunol. Res.*, **3**, 228–235.

Iwashiro,M. *et al.* (1993) Multiplicity of virus-encoded helper T-cell epitopes expressed on FBL-3 tumor cells. *J. Virol.*, **67**, 4533–4542.

Kedzierska,K. *et al.* (2008) Tracking phenotypically and functionally distinct T cell subsets via T cell repertoire diversity. *Mol. Immunol.*, **45**, 607–618.

Laydon,D.J. *et al.* (2015) Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **370**.

Lefranc,M.P. *et al.* (1999) IMGT, the international ImMunoGeneTics database. *Nucl. Acids Res.*, **27**, 209–212.

Levenshtein,V.I. (1966) Binary codes capable of correcting deletions, insertions and reversals. *Soviet Phys. Doklady*, **10**, 707.

Madi,A. *et al.* (2014) T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.*, **24**, 1603–1612.

Mamedov,I.Z. *et al.* (2011) Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol. Med.*, **3**, 201–207.

Mamedov,I.Z. *et al.* (2013) Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front Immunol.*, **4**, 456.

Menezes,J.S. *et al.* (2007) A public T cell clonotype within a heterogeneous autoreactive repertoire is dominant in driving EAE. *J. Clin. Invest.*, **117**, 2176–2185.

Miconnet,I. *et al.* (2011) Large TCR diversity of virus-specific CD8 T cells provides the mechanistic basis for massive TCR renewal after antigen exposure. *J. Immunol.*, **186**, 7039–7049.

Miles,J.J. *et al.* (2010) Genetic and structural basis for selection of a ubiquitous T cell receptor deployed in Epstein-Barr virus infection. *PLoS Pathog.*, **6**, e1001198.

Robins,H.S. *et al.* (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci. Transl. Med.*, **2**, 47ra64.

Serana,F. *et al.* (2009) Identification of a public CDR3 motif and a biased utilization of T-cell receptor V beta and J beta chains in HLA-A2/ Melan-A-specific T-cell clonotypes of melanoma patients. *J. Transl. Med.*, **7**, 21.

Shifrut,E. *et al.* (2013) T cell-receptor repertoire diversity is compromised in the spleen but not in the bone marrow of aged mice due to private and sporadic clonal expansions. *Front. Immunol.*, **CD4**, 4–379.

Shilyansky,J. *et al.* (1994) T-cell receptor usage by melanoma-specific clonal and highly oligoclonal tumor-infiltrating lymphocyte lines. *Proc. Natl. Acad. Sci. U S A*, **91**, 2829–2833.

Thapa,D.R. *et al.* (2015) Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing. *Arthr. Res. Ther.*, **17**, 132.

Turner,S.J. *et al.* (2006) Structural determinants of T-cell receptor bias in immunity. *Nat. Rev. Immunol.*, **6**, 883–894.

UniProt, C (2015) UniProt: a hub for protein information. *Nucl. Acids Res.*, **43**, D204–D212.

Venturi,V. *et al.* (2008) The molecular basis for public T-cell responses? *Nat. Rev. Immunol.*, **8**, 231–238.

Vita,R. *et al.* (2015) The immune epitope database (IEDB) 3.0. *Nucl. Acids Res.*, **43**, D405–D412.