PIRD: Pan immune repertoire database

Wei Zhang[1,2,†], Longlong Wang[1,2,4,†], Ke Liu[1,2,†], Xiaofeng Wei[1,2,†], Kai Yang[1,2], Wensi Du[1,2], Shiyu Wang[1,2,4], Nannan Guo[1,2], Chuanchuan Ma[1,2], Lihua Luo[1,2], Jinghua Wu[1,2], Liya Lin[1,2], Fan Yang[1,2], Fei Gao[1,2], Xie Wang[1,2], Tao Li[1,2], Ruifang Zhang[1,2], Nitin K. Saksena[1,2], Huanming Yang[1,3], Jian Wang[1,3], Lin Fang[1,5], Yong Hou[1,2], Xun Xu[1,2], Xiao Liu[1,2,*]

[1]BGI-Shenzhen, Shenzhen, 518083, China

[2]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

[3]James D. Watson Institute of Genome Sciences, Hangzhou 310058, China

[4]BGI-Education Center, University of Chinese Academy of Sciences, Shenzhen, China.

[5]Department of Biology, University of Copenhagen, Copenhagen, Denmark

* To whom correspondence should be addressed. Tel: 086-13438700710; Fax:; Email: liuxiao@genomics.cn

[†]These authors contributed equally to the paper as first authors.

ABSTRACT

The adaptive immunity is highly specific and mainly comprised of humoral immunity, mediated by antibodies produced by B lymphocytes, and cell-mediated immunity, mediated by T lymphocytes. The huge variety of T and B cell receptor (TCR and BCR) repertoire is vital in directly binding to various auto-and- external antigens. This repertoire is a critical resource for a clear understanding of the immune response and highlights its utility in clinical applications. In recent years, even thousands of samples have been captured for studying TCR and BCR repertoire using sequencing platforms, very few databases have been constructed to store these sequences. To resolve this issue, we have developed a database Pan immune repertoire database (PIRD), located the infrastructure of in China National GeneBank (CNGB) to collect and store the annotated TCR and BCR sequencing data, including Homo sapiens and other species. Except for the primary data storage, PIRD also provides the data visualization function in addition to the interactive online analysis. Additionally, a manually curated database of T- and B-cell receptors targeting known antigens (TBAdb) are also deposited in PIRD. The database is displayed in both English and Chinese on the PIRD website. PIRD can be accessed at https://db.cngb.org/

INTRODUCTION

The rapid sensing and elimination of pathogens is provided by the innate immune system, whereas the adaptive immune system provides a broader and fine-tuned repertoire of recognition for both auto-and foreign antigens. This process involves a tight regulation of antigen-presenting cells and T and B lymphocytes thereby facilitating highly specific immune effector pathways, immunologic memory, and host immune homeostasis regulation. In addition to being highly specific, the adaptive immunity also displays the generation of immunologic memory as one of its key features. As already known that in the event of first encounter with a pathogen the long-lived memory T and B cells are quickly established, and this serves as a protection against the same pathogen (antigen) in the event of second encounter where there is a quick activation of memory cells in mounting a robust and specific immune response against the pathogen(1). The TCR and BCR undergo specific clonal expansion or

antibody affinity improvement in order to generate specific adaptive immune responses. The majority T cells (~95%) are alpha beta T cells (αβ) that α and β chains are used to form the cell receptor, while the small minority of T cells are gamma delta T cells (γδ). BCRs consist of heavy and light chains. During the development of T and B lymphocyte, TCR and BCR are formed by the rearrangement of V, D, J gene segments (a range of V, D and J genes segments located at chromosome), simultaneously, some nucleotides are deleted randomly at the end of gene segments and some nucleotides are inserted during the V-D or D-J junctions (V-J junction for α, γ and light chains)(2). Each lymphocyte carries out the processing of VDJ rearrangement independently. As a consequence, a highly diverse array of TCR and BCR repertoires are generated for each individual, and this diversity in T and B cell repertoire play a crucial role in mounting a specific, robust and sustained response to a foreign pathogen. Theoretically, for human beings, it has been predicted that there would be more than 10^{18} unique TCRs and more than 10^{13} unique BCRs (3,4). Therefore, the extreme large number of receptors produce the individual's strong immune system to respond the complicated and various cellular and external circumstances.

In the last decade, with the advance of sequencing platforms, there has been a rapid accumulation of high-throughput sequencing of TCR and BCR repertoire data. Specifically, since several influential and seminal studies published in 2009(5-7), more and more studies and applications pertinent to this area were published in the subsequent years which included (i) investigation of tumor immune microenvironment, searching tumor-infiltrating T cells targeting neo-antigens and assisting immune therapy(8-12); (ii) detecting minimal residual disease and monitoring of immune reconstitution for leukemia and transplant(13,14); (iii) exploring the immune characteristics of specific diseases (especially in autoimmune and infectious diseases) and evaluating vaccines(15-19); (iv) identifying disease associated clones(20-22), and (v) screening the production of monoclonal antibodies targeting specific antigens and HIV neutralizing antibody(23,24).

It needs to be emphasized that the TCR and BCR sequencing data is a valuable and a very useful resource for other investigators to reuse and search for relevant sequences guiding the processes and mechanisms in adaptive immunity. In addition, such data are extremely useful in search for receptors targeting specific antigens in cancer, neurodegenerative diseases, and viral infections. Therefore, a special curated database to collect and store these sequencing data is highly needed. Adaptive Immune Receptor Repertoire (AIRR) Community have been preparing a database(25), but it might still require considerable time to finish. Recently, the chair of the AIRR Community Executive Felix Breden also published a platform iReceptor for querying and analyzing immune repertoire data(26). Here, we developed a special and multifunctional database PIRD, which can be used for storing the TCR and BCR sequencing data and its corresponding projects, and sample information, in addition to providing a robust platform for users to simultaneously analyse and visualize immune repertoire diversity. The

TCR and BCR sequences are stored with our defined immune repertoire format (IRF). A database of disease and antigen associated sequences TBAdb also deposited in PIRD. The PIRD provides a valuable opportunity for other investigators to reuse the data more easily and to cross sectionally compare against the mined data on TCRs or BCRs. It is envisaged that within the PIRD the data will be regularly updated with newly released TCR and BCR datasets, thereby facilitating a clean and integrated data mining for an improved understanding of immune repertoires needed in clinical applications, future drug and vaccine development.

MATERIAL AND METHODS

Data sources

To collect or upload the data in PIRD, three datasets are required, that include the project information, sample information and annotated TCR or BCR repertoire sequences. The TCR or BCR repertoire can be captured for single chain by multiple PCR(6,27), Rapid Amplification of cDNA Ends (5'RACE)(28), unique molecular identifiers (UMI) (29-31) and other methods. Single-cell RNA sequencing technology(9) and other developed methods(32-34) can obtain paired chains of receptors, such as TCR α and β chains and BCR heavy and light chains. These repertoires are then sequenced by high throughput sequencer, such as Illumina(12), BGISEQ-500(35) and 454(36). The sequences can be processed by several tools, such as MiXCR(37), MiTCR(38), IMonitor(39), IgBLAST(40) and so on, to generate the annotated information. The receptors can be derived from gDNA or RNA sample, or from the peripheral blood cells and specific tissues. Generally, the data in the PIRD are stored independently using consented research studies that have been previously published or will be published in the future. Thus, a detailed information of projects, samples and sequences included in the study is recommended for submission into PIRD. TBAdb, a relatively independent dataset in PIRD, contains the antigen and disease-associated TCR and BCR sequences that were derived from the previously published literatures which were collected and introduced manually into TBAdb.

Database implementation

The implementation of PIRD as a web application and database was performed in Django, a high-level Python Web framework. The framework of search engine and data flows of PIRD, using Elasticsearch (a distributed data and real-time search engine), provide association analysis and real-time and seamless retrieval for all data into the database. The development of online analysis and visualization platform based on the framework of MongoDB. However, all depends on the nodes built by MongoDB clusters and Elasticsearch clusters that standardize and integrate data from them.

RESULTS

Basic datasets in database

The data shown in PIRD are based on the projects so that the users can find detailed information data through the project window or the search section on the website. However, the PIRD mainly includes five basic datasets in the underlying database, including project information, sample information, raw sequencing data, annotated TCR or BCR repertoires, and TBAdb (Fig. 1). The project information contains the description of the project (project name, sample size, types of receptor, species, related published paper and so on) and individuals. One individual's sequences can be from multiple tissues sources(11), multiple timepoints such as vaccine evaluation (30) or longitudinal study for pre-and post-treatment(41,42). Thus, in the interest of accuracy, it is imperative to curate relationship between individual and multiple samples which is a critical element of the project information displayed in PIRD. The sample information includes two types of data. One is the experimental details of samples and clinical indexes, such as sequence type, experimental methods, cell origin, locus, sequencing platform, gender, age and race etc. The other is the overall statistics of sample from the TCR or BCR sequencing repertoires, which include total sequence number, unique Complementarity-determining region 3 (CDR3) number, diversity index (Shannon index), detailed top 10 clone frequencies, CDR3 length distribution, V gene usage, J gene usage and V-J pairings. Most of the statistics provide figures to users for direct visualisation. Raw sequencing data are deposited in the Nucleotide Sequence Archive (CNSA, https://db.cngb.org/cnsa/) that is a new database at the CNGB, storing different kinds of raw sequencing data. However, the raw data information is described in the PIRD, and the raw data for users visiting the website can be retrieved by clicking on the weblink. The dataset of annotated TCR or BCR repertoire is the core of this database. We have defined a special format to store these data, called immune repertoire format (IRF). The different CDR and Framework region (FR) fragments, VDJ assignment and paired chains (such as alpha and beta) are shown in it. The annotation is displayed for each sequence, but all of the sequences from one sample are, as a whole, are meant for uploading and analysis. Lastly, there is a special dataset TBAdb deposited in the PIRD. The data of TBAdb are the antigens associated TCRs or BCRs which have been reported in previously published papers. Up to now, >70,000 sequences associated with 94 different diseases are embodied in TRAdb. Additionally, in the PIRD website, we also provide Tools and Document window to deposit new softwares related to immune repertoire analysis and some format specifications and user manuals which will assist the end users of the PIRD.

Functions of database

Overall, PIRD provides five functions, including data storage, data statistics/visualization, interactive online analysis/visualization, search, and uploading/downloading of data (Fig. 1). Firstly, like other databases, data storage is the primary functional objective of PIRD. With the exception of raw sequence data, other basic datasets are directly stored in this database. Secondly, for each sample, the total sequence number, CDR3 diversity, unique CDR3 number, CDR3 length distribution, V/J/V-J pairing usage etc., are calculated from the annotated TCR or BCR repertoires (Fig. 1). These sample statistics reflect the diversity of immune repertoire, which are commonly used in evaluating the disease status(42) and compare with different diseases or phenotypes(43). To facilitate users to visualize the statistical data, the PIRD provides corresponding figures of statistical evaluation with a single click on the weblink. As an example, Fig. 2A shows the one sample's V-J gene pairing and CDR3 length distribution. Additionally, we also provide the visual figures for the clinical and other data pertinent to the projects. Through these figures, one can directly obtain the overall information of samples related to the project, such as the age distribution and CDR3 diversity distribution etc. (Fig. 2B). The third function of this database is the online interactive analysis, which is the main difference that demarcates PIRD from other commonly used databases. We offer a simple and a user-friendly platform for all users, including those who lack the background bioinformatics, to compare and visualise datasets. Users can upload the data themselves, or just select partial PIRD data to do selective analysis and visualization. In this database, we provide four types of search modules, including search by project, sample, clone, TBAdb and basic local alignment (BLAST) (Fig. 2C). The first three types contain multiple but restricted items for users to choose and search the data of interest. The fourth type searching by TRAdb is a special search window dependent on the dataset of TBAdb, including antigens associated TCR αβ, TCR γδ and BCR sequences. The last type is a fuzzy search that finds regions of local similarity between sequences using BLAST tool(44,45). This compares query to sequences in database and calculates the statistical significance of matches. Using the BLAST search, the website goes into an independent interface with five BLAST methods (BLASTN, BLAST, BLASTX, TBLASTN, TBLASTX) which can be used for further analysis (Fig. 2C). Lastly, the database offers data upload and download function. For uploading the data, project information, sample information (experimental and clinical data) and annotated TCR or BCR repertoire are recommended for the data to be included.

Online analysis platform in the database

To facilitate users to understand and use the data in PIRD, we have developed the platform of interactive online analysis, and some simple analyses such as the comparison of different groups' datasets can performed and the corresponding visual data in form of figures can be obtained as an output. Users can upload the data themselves or just select partial data of interest in PIRD, and then a datasheet will be created to merge the selected data (Fig. 3A). The analyses include multiple items'

comparison of either several samples or multiple groups by dragging the button on the left of the website page. We offer a choice of three types of figures to show the results, which include histogram plot, line plot and box plot. As an example, in the PIRD, we use the annotated data from a patient from breast cancer project to show (Fig. 3B). We find that the data from this patient display inconsistent CDR3 length and TCR Vβ gene usage distributions among tumour, adjacent normal tissue and lymph node. These results imply that the infiltrating T cells in the tumour are significantly different from others. To show the function of comparison for multiple groups, we selected data of colon cancer project, breast cancer project, healthy people project and systemic lupus erythematosus (SLE) project from PIRD (Fig. 3C). From the results of colon cancer data, we can find the diversity of intestinal segment C1 was lower than other segments, and the difference in genetic diversity between tumour and the adjacent normal tissues. However, the greatest advantage of analysis platform is to compare multiple diseases deposited in the PIRD, which cannot be done in a single study. The comparison of TCR repertoire for health, SLE and breast cancer showed that the diversity of SLE was lower than others, which might be attributed to the T cells proliferated in response to the autoantigens (Fig. 3C, right panel). There was also a significant distinction between TCR β repertoire among tumor, adjacent normal tissue, lymph node and peripheral blood mononuclear cell (PBMC) for breast cancer patients (Fig. 3C, right panel).

TBAdb: a manually curated database of T- and B-cell receptors targeting known antigens

Antigen specific TCRs or BCRs are critical resource to understand the adaptive immune system and have a significant potential in translational medicine, such as disease assessment, therapy and vaccine development. These data can be used in annotating the large-scale of TCR or BCR repertoire by high throughput sequencing. To collect previously reported disease and antigens-associated sequences, we read hundreds of published literatures and created a knowledge database of the extracted TCR and BCR sequences with relevant information. All the relevant information collected for each sequence were manually curated and critically and unambiguously checked by two people. The fields include the disease name, antigen, CDR3 sequence, V/J usage, HLA type, experimental method, grade, publication details, and so on. Importantly, the grade for each sequence is provided to evaluate the quality and reliability of relative antigen, according to the method identifying antigen-specific sequences. All fields, including the criterion of grade, are described in detail in the supplementary file and in Tools and Docs section in database website. Recently, two database VDJdb(46) and McPAS(47) have published and both have collected abundant TCR sequences associated with multiple disease. We also considered the sequences in VDJdb and McPAC. All sequences, including our collected novel sequences and others, integrated into the same format and stored in the database of TBAdb. It contains three parts, including TCR αβ, TCR γδ and BCR. TBAdb currently includes

78,202 sequences, encompassing a total of 94 different diseases. We classified these diseases into five categories that include- autoimmunity, cancer, pathogen, allergy and other. Each of the first three disease categories contains more than 20 diseases (Fig. 4A). In TBAdb, the pathogen category accounts for most of the sequences, which reaches to 75.91% of the total sequences (Fig. 4B). We also found that most of the sequences in diseases of pathogen category are more reliable with grade >4 (Fig. 4C). This could be because the pathogen diseases have been widely researched and the TCR or BCR sequences targeting pathogen are relatively easy to identify. The second largest proportion of category is the cancer (Fig. 4B), but more than half of sequences are showed with lower grade (Fig. 4C). Although the other category contains 20 diseases, it accounts for only 3.5% of sequences and most of them are less reliable (Fig. 4C).

IRF format instruction

Although a large number of samples have been sequenced for studying the TCR and BCR repertoire, there is no general format to store these data, which results in inconvenience for people using these databases, communicate and exchange the dataset. To resolve this issue, we propose a text-based format IRF (V1.0, Supplementary file) for immune repertoire sequencing data. In the format, each line typically represents the annotated information of unique nucleotide sequence. The annotated information includes the alignment records, VDJ assignment, deleted and inserted nucleotides, and structure of clone such as the details of CDRs and FRs. If paired two chains, such as TCR α and β or heavy and light chains, are sequenced at the same time, the paired sequences can be deposited at the adjacent two lines, which can be identified by the same ID. Each line has 27 mandatory fields, which include almost all of information used for immune repertoire analysis. These fields always appear in the same order and must be present, but their values can be used “na” instead if the corresponding information is unavailable. All fields are described in details in supplementary file and in Tools and Docs section of PIRD website. The immune repertoire sequencing data can be processed using several software, such as MiXCR(37), MiTCR(38), IMonitor(39), IgBLAST(40), Decombinator(48), and results from these tools could be easy to transform into IRF format.

DISCUSSION

Here, we describe the Pan Immune Repertoire Database (PIRD), a unified repository to facilitate scalable data mining of immune repertoires in both nucleotide and amino acid forms. It is the lack of well-defined and annotated repositories required for the immune repertoire sequence storage prompted the creation of PIRD. The PIRD is located at the CNGB, Shenzhen, China, and is a newly developing integrative and interactive database. It includes more than 20 biologically diverse disease

databases, along with the integration of the deposited raw sequencing database CNSA, and so on. From the search page of CNGB, the sequences in PIRD can be find out if there are similar with the query sequence. Starting from the first version of release in 2016, PIRD has collected more than 1800 samples with the large-scale of TCR and BCR repertoire sequencing data, including Homo sapiens and primate species. This multifunctional database brings several benefits to users. Firstly, the high diversity of immune repertoire and few clones overlapping between individuals(49), result in the fact that large-scale samples are required to evaluate and validate the reliability of each study. Identifying the disease or antigen-specific clones is the main purpose of some studies. If the identified clones can be searched and compared with the various other disease data in PIRD, and if only few samples can find these clones in PIRD, it can bring a strong validation for these disease-associated clones and possibly disease outcomes. Secondly, the bioinformatic methods for immune repertoire data are still in its infancy. Some clustering methods of TCR sequences targeting the same antigen have been reported recently(21,22), and more machine learning methods would be developed in the future due to the increase in the volume of the sequencing data. Thus, the project data in PIRD can be reused and mined for some in-depth analyses using new methods, and multiple diseases in database can be compared simultaneously and seamlessly. Thirdly, the online analysis function of the database offers a valuable platform for users to analyze the data of interest with a simple interactive interface. Lastly, the integrated database of TBAdb provides a vital resource to annotate the large-scale immune repertoire sequencing data.

The future development of PIRD will further integrate and optimize the platform of online analysis, which will provide a more friendly interface for users to operate with more integrative functions and web-based tools. We hope PIRD will act as a multifunction database, which is not only meant for data storage but also for visualization and analysis platform. On the other hand, we will continue to collect more sequencing data to store in PIRD in the future and facilitate the evolution of PIRD into the largest integrative and interactive database for immune repertoire studies, that will include the previously published and unpublished datasets from highly reputable reliable sources. Additionally, we will continue to update the data of TBAdb in real time. Most of sequences in currently TBAdb are TCR αβ sequences, and so in the near future, sequences of BCR associated with various antigens and diseases will also be collected and displayed.

We believe that this database should aid in-depth comparative and cross-sectional analyses encompassing different disease and immunologic states to discern difference and the commonalities between immune repertoires. Moreover, such comparative studies of immune repertoires using PIRD will pave the way in for engineering better biotherapeutics, vaccines and immune therapies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

This project was funded by the Beijing Genomics Institute and China National GeneBank. The funder provided support in the form of salaries of all authors.

FUNDING

This work was supported in part by Shenzhen Municipal Government of China (JCYJ20160510141910129).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

TABLE AND FIGURES LEGENDS

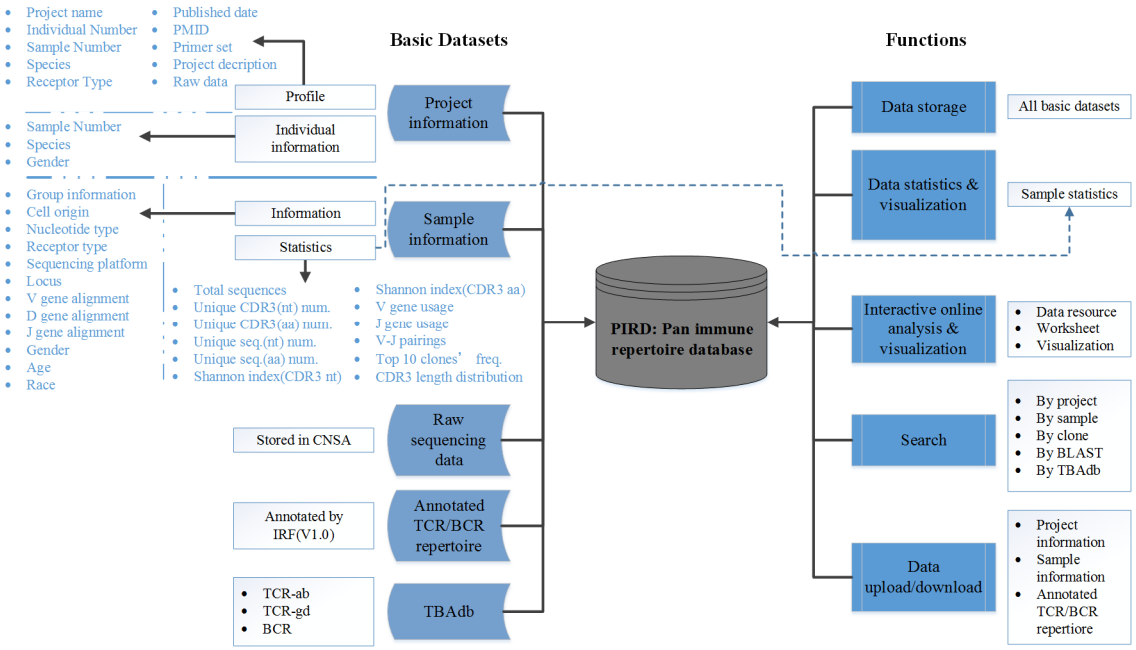

Figure 1. Systematic design of the PIR database. In this database, there are five basic datasets with five different functions. Each dataset or function includes one or multiple parts, and most parts contain several items. Part of these items are listed on the left, and some on the right side of the schema shown above.

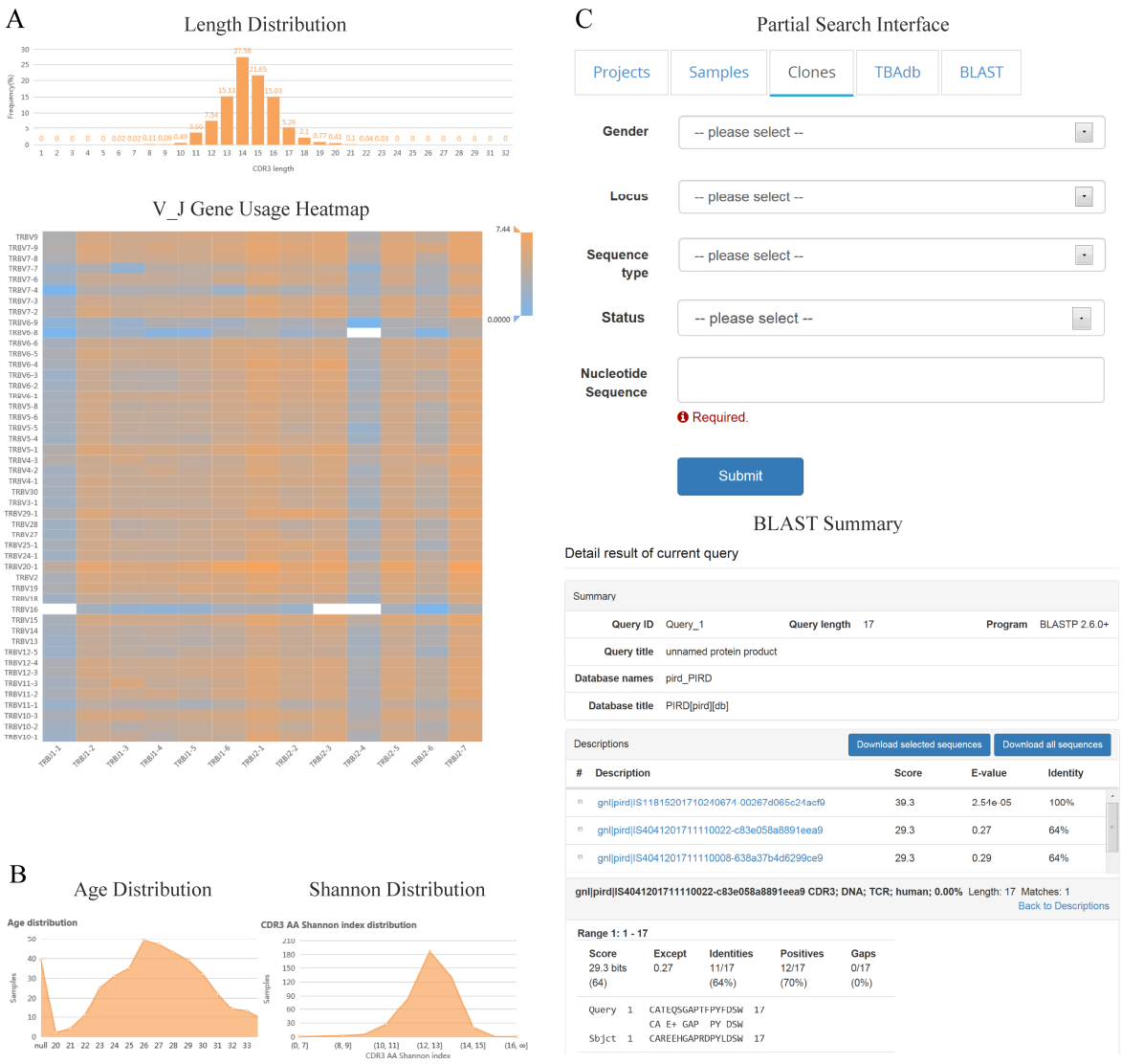

Figure 2. Examples for visualization and search functions. (A). Visualization of the project information. The left panel shows the age distribution of samples in a project, and the right panel shows the CDR3 diversity distribution for a project. (B). Visualization of statistics for a given sample. The proportion of each V-J gene pairing in a sample is represented in colour (top panel). The length distribution of CDR3 amino acid sequence for a sample is displayed in the bottom panel. (C). Partial search interface (top panel) and example by local alignment search (bottom panel). The search by

project, sample, clone and TBAdb are provided for multiple choices to restrict the range of content. The local alignment search uses the tool BLAST to find sequences analogous to the query sequences.

Figure 3. Examples of online interactive analysis and visualization. (A). flowchart of analysis platform. (B). The length distribution of CDR3 amino acid sequence and TCR Vβ gene usage from one patient of breast cancer. There were three samples from the patient, including tumor, adjacent normal tissue and lymph node. The data were selected from breast cancer project in PIRD. (C) The diversity of CDR3 amino acid sequences among different groups. The left panel shows the BCR diversities of intestinal samples among healthy, adenoma and carcinoma groups from colon cancer project in the PIRD. Each healthy individual contains seven intestinal segments (C1, C2, C3, C4, C5, C6 and C7). The right panel shows the TCR repertoire's diversities from health, SLE and breast cancer projects in PIRD. The breast cancer project contains four types of samples, including tumor, adjacent normal tissue, lymph node and PBMC. The diversity is represented by Shannon index. All results of (B) and (C) are done on the analysis platform in PIRD.

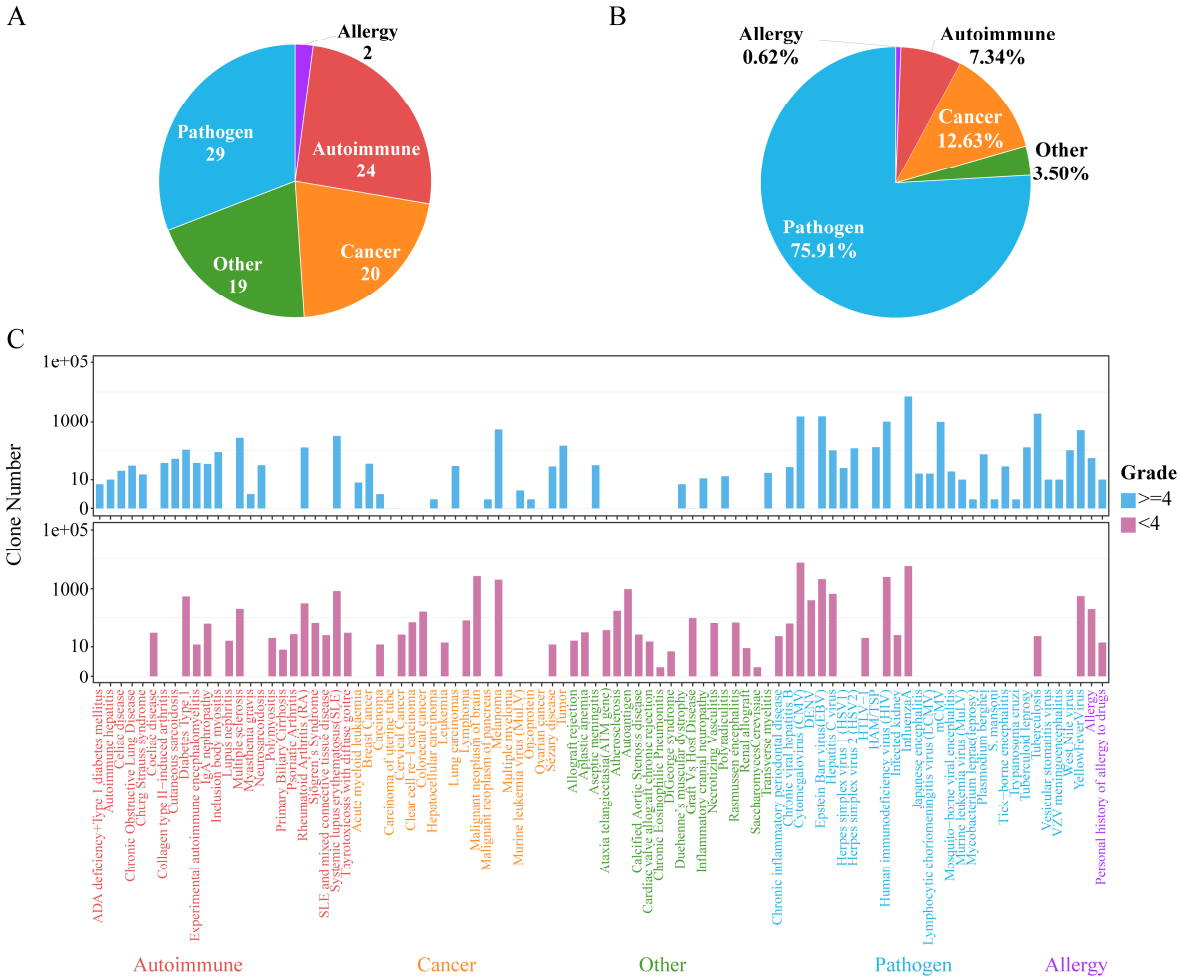

Figure 4. overall summary of antigen associated sequences in TBAdb. (A). The number of diseases in five disease categories. (B). the proportion of sequences in each disease categories. (C). The distribution of sequences number in each disease in TBAdb. The upper panel shows the sequences with the grade more than 4, and the bottom panel shows the sequences with the grade less than 4. The grade is a score to evaluate the reliability of sequences associated with the antigen ranging from 1 to 9. Other, signifies the disease that could not be classified into the four disease categories.

REFERENCES

1. Cooper, M.D. and Alder, M.N. (2006) The evolution of adaptive immune systems. *Cell*, **124**, 815-822.

2. Liu, X. and Wu, J. (2018) History, applications, and challenges of immune repertoire research. *Cell biology and toxicology*.

3. Venturi, V., Price, D.A., Douek, D.C. and Davenport, M.P. (2008) The molecular basis for public T-cell responses? *Nature reviews. Immunology*, **8**, 231-238.

4. Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H. and Quake, S.R. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, **32**, 158-168.

5. Weinstein, J.A., Jiang, N., White, R.A., 3rd, Fisher, D.S. and Quake, S.R. (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324**, 807-810.

6. Robins, H.S., Campregher, P.V., Srivastava, S.K., Wacher, A., Turtle, C.J., Kahsai, O., Riddell, S.R., Warren, E.H. and Carlson, C.S. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, **114**, 4099-4107.

7. Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., Hanczaruk, B., Nguyen, K.D. *et al.* (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine*, **1**, 12ra23.

8. Reuben, A., Gittelman, R., Gao, J., Zhang, J., Yusko, E.C., Wu, C.J., Emerson, R., Zhang, J., Tipton, C., Li, J. *et al.* (2017) TCR Repertoire Intratumor Heterogeneity in Localized Lung Adenocarcinomas: An Association with Predicted Neoantigen Heterogeneity and Postsurgical Recurrence. *Cancer discovery*, **7**, 1088-1097.

9. Zheng, C., Zheng, L., Yoo, J.K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J.Y., Zhang, Q. *et al.* (2017) Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell*, **169**, 1342-1356 e1316.

10. Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., Sims, J.S., Hodi, F.S., Martin-Algarra, S., Mandal, R., Sharfman, W.H. *et al.* (2017) Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, **171**, 934-949 e915.

11. Wang, T., Wang, C., Wu, J., He, C., Zhang, W., Liu, J., Zhang, R., Lv, Y., Li, Y., Zeng, X. *et al.* (2017) The Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues. *Cancer immunology research*, **5**, 148-156.

12. Zhang, W., Feng, Q., Wang, C., Zeng, X., Du, Y., Lin, L., Wu, J., Fu, L., Yang, K., Xu, X. *et al.* (2017) Characterization of the B Cell Receptor Repertoire in the Intestinal Mucosa and of Tumor-Infiltrating Lymphocytes in Colorectal Adenoma and Carcinoma. *Journal of immunology*, **198**, 3719-3728.

13. Wu, D., Sherwood, A., Fromm, J.R., Winter, S.S., Dunsmore, K.P., Loh, M.L., Greisman, H.A., Sabath, D.E., Wood, B.L. and Robins, H. (2012) High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Science translational medicine*, **4**, 134ra163.

14. Wu, J., Jia, S., Wang, C., Zhang, W., Liu, S., Zeng, X., Mai, H., Yuan, X., Du, Y., Wang, X. *et al.* (2016) Minimal Residual Disease Detection and Evolved IGH Clones Analysis in Acute B Lymphoblastic Leukemia Using IGH Deep Sequencing. *Frontiers in immunology*, **7**, 403.

15. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A. and Peakman, M. (2017) T cell receptor beta-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nature communications*, **8**, 1792.

16. Smithey, M.J., Venturi, V., Davenport, M.P., Buntzman, A.S., Vincent, B.G., Frelinger, J.A. and Nikolich-Zugich, J. (2018) Lifelong CMV infection improves immune defense in old mice by broadening the mobilized TCR repertoire against third-party infection. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, E6817-E6825.

17. Parameswaran, P., Liu, Y., Roskin, K.M., Jackson, K.K., Dixit, V.P., Lee, J.Y., Artiles, K.L., Zompi, S., Vargas, M.J., Simen, B.B. *et al.* (2013) Convergent antibody signatures in human dengue. *Cell host & microbe*, **13**, 691-700.

18. Jiang, N., He, J., Weinstein, J.A., Penland, L., Sasaki, S., He, X.S., Dekker, C.L., Zheng, N.Y., Huang, M., Sullivan, M. *et al.* (2013) Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine*, **5**, 171ra119.

19. Jackson, K.J., Liu, Y., Roskin, K.M., Glanville, J., Hoh, R.A., Seo, K., Marshall, E.L., Gurley, T.C., Moody, M.A., Haynes, B.F. *et al.* (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell host & microbe*, **16**, 105-114.

20. Emerson, R.O., DeWitt, W.S., Vignali, M., Gravley, J., Hu, J.K., Osborne, E.J., Desmarais, C., Klinger, M., Carlson, C.S., Hansen, J.A. *et al.* (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature genetics*, **49**, 659-665.

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. 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.

23. Jardine, J.G., Kulp, D.W., Havenar-Daughton, C., Sarkar, A., Briney, B., Sok, D., Sesterhenn, F., Ereno-Orbea, J., Kalyuzhniy, O., Deresa, I. *et al.* (2016) HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science*, **351**, 1458-1463.

24. Cheung, W.C., Beausoleil, S.A., Zhang, X., Sato, S., Schieferl, S.M., Wieler, J.S., Beaudet, J.G., Ramenani, R.K., Popova, L., Comb, M.J. *et al.* (2012) A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature biotechnology*, **30**, 447-452.

25. Rubelt, F., Busse, C.E., Bukhari, S.A.C., Burckert, J.P., Mariotti-Ferrandiz, E., Cowell, L.G., Watson, C.T., Marthandan, N., Faison, W.J., Hershberg, U. *et al.* (2017) Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature immunology*, **18**, 1274-1278.

26. Corrie, B.D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F.M.W., Christley, S., Scott, J.K. *et al.* (2018) iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological reviews*, **284**, 24-41.

27. Liu, X., Zhang, W., Zeng, X., Zhang, R., Du, Y., Hong, X., Cao, H., Su, Z., Wang, C., Wu, J. *et al.* (2016) Systematic Comparative Evaluation of Methods for Investigating the TCRbeta Repertoire. *PloS one*, **11**, e0152464.

28. Warren, R.L., Freeman, J.D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J.R. and Holt, R.A. (2011) Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome research*, **21**, 790-797.

29. Turchaninova, M.A., Davydov, A., Britanova, O.V., Shugay, M., Bikos, V., Egorov, E.S., Kirgizova, V.I., Merzlyak, E.M., Staroverov, D.B., Bolotin, D.A. *et al.* (2016) High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nature protocols*, **11**, 1599-1616.

30. Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L. and Quake, S.R. (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 13463-13468.

31. Khan, T.A., Friedensohn, S., Gorter de Vries, A.R., Straszewski, J., Ruscheweyh, H.J. and Reddy, S.T. (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Science advances*, **2**, e1501371.

32. Howie, B., Sherwood, A.M., Berkebile, A.D., Berka, J., Emerson, R.O., Williamson, D.W., Kirsch, I., Vignali, M., Rieder, M.J., Carlson, C.S. *et al.* (2015) High-throughput pairing of T cell receptor alpha and beta sequences. *Science translational medicine*, **7**, 301ra131.

33. DeKosky, B.J., Kojima, T., Rodin, A., Charab, W., Ippolito, G.C., Ellington, A.D. and Georgiou, G. (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature medicine*, **21**, 86-91.

34. DeKosky, B.J., Ippolito, G.C., Deschner, R.P., Lavinder, J.J., Wine, Y., Rawlings, B.M., Varadarajan, N., Giesecke, C., Dorner, T., Andrews, S.F. *et al.* (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology*, **31**, 166-169.

35. Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., Qu, S., Mei, X., Chen, H., Yu, T. *et al.* (2017) A reference human genome dataset of the BGISEQ-500 sequencer. *GigaScience*, **6**, 1-9.

36. Wang, C., Liu, Y., Cavanagh, M.M., Le Saux, S., Qi, Q., Roskin, K.M., Looney, T.J., Lee, J.Y., Dixit, V., Dekker, C.L. *et al.* (2015) B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 500-505.

37. Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V. and Chudakov, D.M. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nature methods*, **12**, 380-381.

38. Bolotin, D.A., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Turchaninova, M.A., Zvyagin, I.V., Britanova, O.V. and Chudakov, D.M. (2013) MiTCR: software for T-cell receptor sequencing data analysis. *Nature methods*, **10**, 813-814.

39. Zhang, W., Du, Y., Su, Z., Wang, C., Zeng, X., Zhang, R., Hong, X., Nie, C., Wu, J., Cao, H. *et al.* (2015) IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. *Genetics*, **201**, 459-472.

40. Ye, J., Ma, N., Madden, T.L. and Ostell, J.M. (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, **41**, W34-40.

41. Vollmers, C., De Vlaminck, I., Valantine, H.A., Penland, L., Luikart, H., Strehl, C., Cohen, G., Khush, K.K. and Quake, S.R. (2015) Monitoring pharmacologically induced immunosuppression by immune repertoire sequencing to detect acute allograft rejection in heart transplant patients: a proof-of-concept diagnostic accuracy study. *PLoS medicine*, **12**, e1001890.

42. Wood, B., Wu, D., Crossley, B., Dai, Y., Williamson, D., Gawad, C., Borowitz, M.J., Devidas, M., Maloney, K.W., Larsen, E. *et al.* (2018) Measurable residual disease detection by high-throughput sequencing improves risk stratification for pediatric B-ALL. *Blood*, **131**, 1350-1359.

43. Liaskou, E., Klemsdal Henriksen, E.K., Holm, K., Kaveh, F., Hamm, D., Fear, J., Viken, M.K., Hov, J.R., Melum, E., Robins, H. *et al.* (2016) High-throughput T-cell receptor sequencing across chronic liver diseases reveals distinct disease-associated repertoires. *Hepatology*, **63**, 1608-1619.

44. Ye, J., McGinnis, S. and Madden, T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic acids research*, **34**, W6-9.

45. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403-410.

46. 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. *et al.* (2018) VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic acids research*, **46**, D419-D427.

47. 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*, **33**, 2924-2929.

48. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. and Chain, B. (2013) Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, **29**, 542-550.

49. Glanville, J., Kuo, T.C., von Budingen, H.C., Guey, L., Berka, J., Sundar, P.D., Huerta, G., Mehta, G.R., Oksenberg, J.R., Hauser, S.L. *et al.* (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 20066-20071.