Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz872

VDJbase: an adaptive immune receptor genotype and haplotype database

Aviv Omer[1,†], Or Shemesh[1,†], Ayelet Peres[1,†], Pazit Polak[1], Adrian J. Shepherd[2], Corey T. Watson[3], Scott D. Boyd[4], Andrew M. Collins[5], William Lees[2] and Gur Yaari[1,*]

[1]Bioengineering, Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel, [2]Institute of Structural and Molecular Biology, Birkbeck, University of London, London, UK, [3]University of Louisville School of Medicine, Biochemistry and Molecular Genetics, Louisville, KY 40292, USA, [4]Department of Pathology, Stanford University, Stanford, CA 94305, USA and [5]School of Biotechnology and Biomolecular Sciences, University of NSW, Kensington, Sydney, NSW 2052, Australia

Received August 11, 2019; Revised September 19, 2019; Editorial Decision September 24, 2019; Accepted October 01, 2019

ABSTRACT

VDJbase is a publicly available database that offers easy searching of data describing the complete sets of gene sequences (genotypes and haplotypes) inferred from adaptive immune receptor repertoire sequencing datasets. VDJbase is designed to act as a resource that will allow the scientific community to explore the genetic variability of the immunoglobulin (Ig) and T cell receptor (TR) gene loci. It can also assist in the investigation of Ig- and TR-related genetic predispositions to diseases. Our database includes web-based query and online tools to assist in visualization and analysis of the genotype and haplotype data. It enables users to detect those alleles and genes that are significantly over-represented in a particular population, in terms of genotype, haplotype and gene expression. The database website can be freely accessed at https://www.vdjbase.org/, and no login is required. The data and code use creative common licenses and are freely downloadable from https://bitbucket.org/account/user/yaarilab/projects/GPHP.

INTRODUCTION

An important application of the recent advances in high throughput DNA sequencing is the exploration of adaptive immune receptor repertoires (AIRR). AIRR sequencing (AIRR-seq) enables exploration of the dynamics of the adaptive immune system (1), and has applications to the study of aging (2,3), cancer (4), autoimmune diseases (5–7), allergy (8), infectious diseases (9) and vaccine design (10).

A crucial step in the analysis of AIRR-seq data is the correct identification of specific V, D and J germline genes that contribute to each antibody and T cell receptor gene sequence. It is the starting point for in-depth analyses such as the identification and quantification of somatic hypermutation (11), determination of gene usage distribution, and correlation of AIRR-seq data with clinical conditions (12). For example, it was recently demonstrated that the presence or absence of a specific allele greatly affects the response to influenza A and HIV infections (13–15). Other infectious diseases as well as cancer and allergy may also be sensitive to the germline repertoire. However, our knowledge of the genetic loci encoding Ig and TR is very incomplete, since the genomic regions encoding these receptors contain many duplications, deletions, and other complex events, which hinder their direct sequencing using short reads (16). This is true for all studied species to date, including humans. Genomic studies of the human loci have come from just a handful of individuals, and we therefore do not know the extent of population variation within these loci, though there is reason to believe the variation is significant (17–21). Recently, we and others have published several computational tools to help explore these regions, to infer previously unknown alleles, deletion polymorphisms, and complete sets of immungolobulin genes that are expressed by different individuals (genotypes and haplotypes) from AIRR-seq data (21–28).

Germline sequences affirmed by the new tools are curated in the international ImMunoGeneTics (IMGT) information system (29) after review by the Inferred Allele Review Committee (IARC) of the AIRR Community (30). This process is facilitated by OGRDB (the Open Germline Receptor Database: https://ogrdb.airr-community.org), which provides supporting evidence for published alleles, including details of repertoires in which they have been observed. Currently, there are ∼60 alleles that are either under review or have recently been affirmed by the IARC and accepted into IMGT. However, data relating to genotypes, haplo-

*To whom correspondence should be addressed. Tel: +972 3 7384625; Email: gur.yaari@biu.ac.il
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

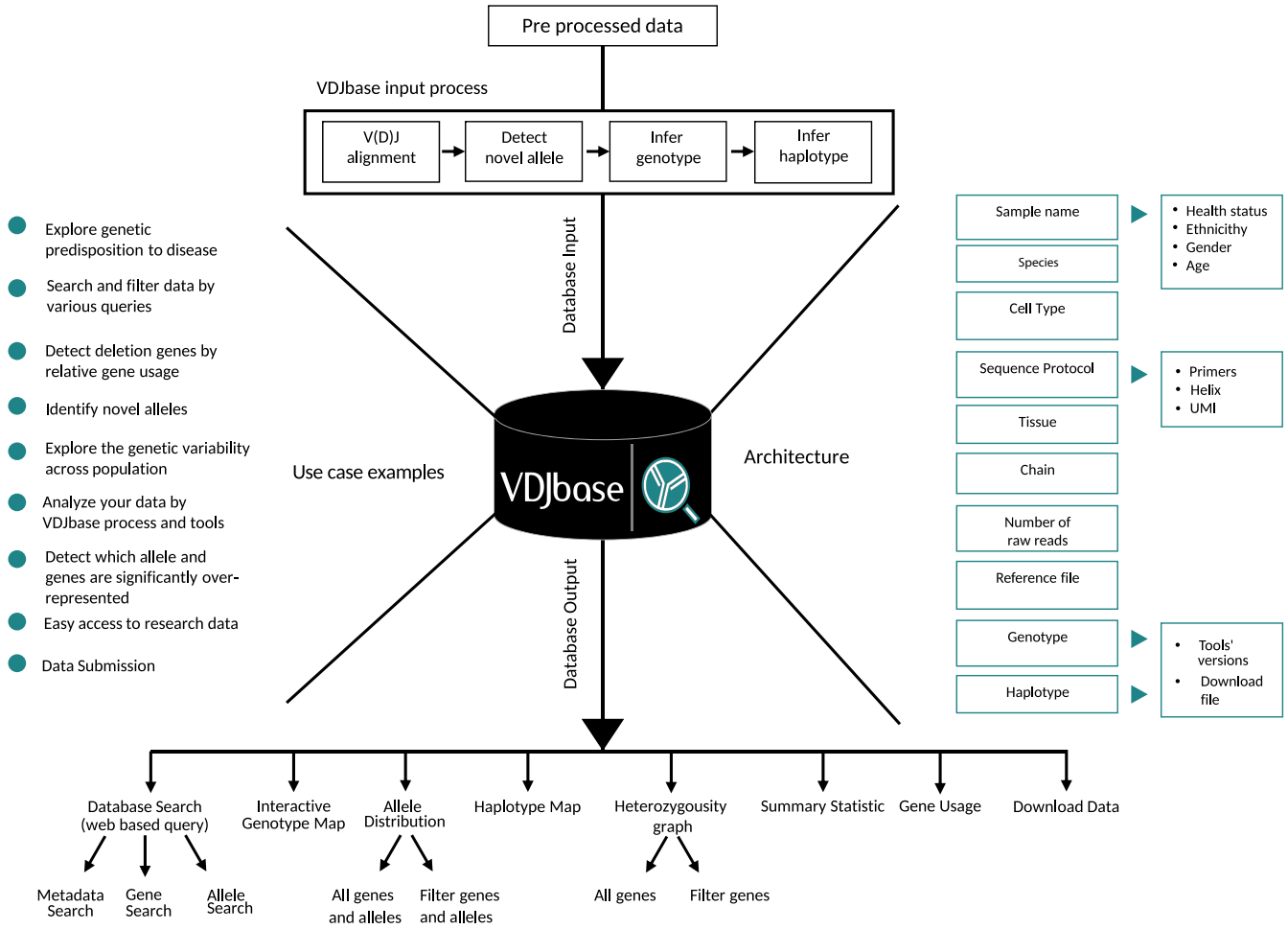

Figure 1. Schematic chart of VDJbase workflow.

types, and general gene usage across the human population are currently beyond the scope of IMGT and OGRDB. There is a need to better understand the usage of germline alleles in different individuals, ethnic and clinical groups. A better picture of the set of alleles expressed by each individual should lead to important discoveries such as predispositions to disease and variable responses to vaccination and drug therapy. Currently, the prevalence within the human population of each allele curated in IMGT is unclear, and the very existence or functionality of many sequences has even been questioned (31). For this reason, we have developed VDJbase, a publicly available database that offers easy searching of antibody genotype and haplotype data inferred from AIRR-seq datasets. VDJbase stores information about genotypes and haplotypes inferred from individuals from diverse ethnic and clinical backgrounds, and produces summary statistics about sets of samples that are filtered according to their associated meta-data.

IMPLEMENTATION

The web interface of VDJbase offers researchers a fast and convenient way to browse for genotypes and haplotypes, compare published datasets, generate interactive visual analyses, and submit AIRR-seq data to foster continuous growth. To allow for unbiased comparisons, the database inputs are generated by an identical data processing pipeline. The pipeline's input is pre-processed Ig and TR sequences. The pipeline begins with a preliminary V(D)J assignment, that includes an inference of previously unknown alleles, followed by inference of genotype and haplotype (see materials and methods section). Users can freely access the database using any browser. Results are displayed as a table, along with samples and their related metadata, providing files and figures that can be downloaded to the user's own computer. We exploit the HTML platform for interactive visualizations. Unlike static charts, interactive data visualizations encourage users to explore and even manipulate the data to uncover other factors. See Figure 1 for a schematic diagram of VDJbase. The entry page includes tutorials about the database service, with links to the user guide and to the content search.

Database search

Browsing is a very useful capability in VDJbase, which can be easily searched by selecting samples of interest and output fields. Users can interactively interrogate genotypes and

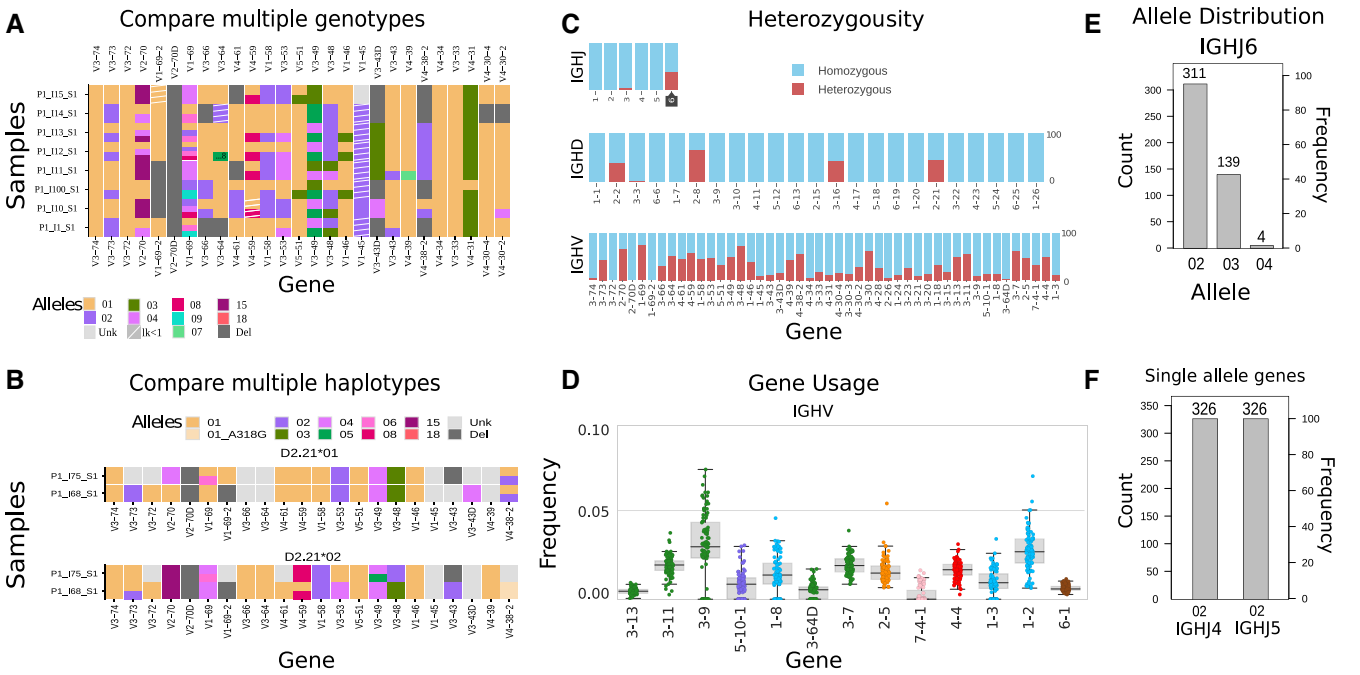

Figure 2. Data visualizations in VDJbase. (**A**) Comparison between 8 genotypes. Row and column represent individual and gene, respectively. Colors correspond to alleles. (**B**) Comparison between two haplotypes. Haplotype is inferred using the IGHD2-21 gene as an anchor gene. The upper panel corresponds to the chromosome carrying IGHD2-21*01 and the lower panel to the chromosome carrying IGHD2-21*02. (**C**) Heterozygosity abundance for each gene in the samples of interest. (**D**) Gene usage. Each point represents an individual. Colors correspond to V family gene. (**E**) Allele appearance of IGHJ6. Left Y axis corresponds to the number of individuals, and the right Y axis corresponds to the frequency of the allele in the samples of interest. (**F**) Allele appearance of IGHJ4 and IGHJ5. Y axis is the same as in (**E**). Gene order in all graphs correspond to their location on the chromosome.

haplotypes using the 'Database Search' page. Searches can be performed by various queries, such as cell type (e.g. memory B cell), tissue type (e.g. blood), health status (e.g. celiac), Ig group (e.g. heavy chain), isotype (e.g. IgM), sex or specific genes and alleles. For a better view of the information, we have established a capability to generate user-friendly visualization graphs. The 'Export Graphs' menu enables the creation of a visual analysis of the user's selected samples. Using a set of drop-down tag lists, users can filter all visualizations according to genes, alleles, or the certainty level of inferences (K_{diff}, see (21,26)) for each genotype/haplotype decision. All entries can be downloaded using the 'Download Selected' tab, which provides a .zip file with the selected data and related meta-data that can be viewed, for instance, in Microsoft Excel. This combination of the annotation and the availability of the underlying data for large sample sets is currently available nowhere else for AIRR-seq data, and is a step forward in the context of open data sharing. Graphs are downloadable in PDF file format. To allow users to quickly assess the complete information on the experimental set-up and materials used for their selected sample, the 'Reference' column contains clickable icons which open any manuscript describing the data in a new browser window. Each section contains help materials to ensure ease of use, without prerequisite knowledge or experience. Clicking the '?' symbol located to the right of each item pops open an explanation. On the 'Explore Data' page, users can view representative examples of interesting findings revealed by VDJbase, together with a summary of the number of studies, samples and related metadata currently stored in the database.

Visualization and analysis

Apart from PDF format, VDJbase uses the Javascript graphing library plotly.js to provide online, interactive data visualization designed to help users gain insights into the data. Genotypes stored in the database include a measure of certainty of the genotype call for each gene (K_{diff}). Briefly, this measure is the ratio between the models' likelihoods calculated from the posterior probability distribution. A 'Genotype' tab creates an interactive graph which allows users to modify parameters to explore the genotype data according to their interests. For instance, users can focus on specific alleles or screen the results by certainty level (K_{diff}). The graph can visualize 1–20 genotypes in a single page, to facilitate comparison between individuals. Comparison of a larger number of genotypes and haplotypes is enabled through a heatmap graph (see Figure 2A and B).

Inference of personal genotypes also allows us to estimate the heterozygosity of genes in the population. We consider genes for which more than one allele is carried by an individual to be heterozygous. The 'Heterozygous' graph allows users to assess the level of hemizygosity/heterozygosity/homozygosity for each gene in different populations. To enable users to obtain more specific information, frequency values and raw counts appear

as pop-ups when users hover above the corresponding bar (Figure 2C).

The 'Gene usage' graph provides a view of gene expression in the population (Figure 2D). Each point in the graph represents a single individual, and colors represent the gene families. The order of genes is based on their chromosomal location (17).

The 'Allele distribution' graph represents the allele distribution within a selected population, for each gene. Users can compare the distributions of alleles between different populations (Figure 2E and F). VDJbase utilizes two R packages for visualizations, 'vdjbaseVis' and 'RAbHIT' (28). These packages can be downloaded and used for free under the CC BY-SA 4.0 license from https://bitbucket.org/account/user/yaarilab/projects/GPHP.

Data submission

To enable users to submit their own published AIRR sequences, a straightforward submission form is available upon request via our 'Discussion Forum'. Submitted forms are validated and the associated data processed by the site administrator. This will allow the repository to grow, and for contributors to receive appropriate credit from database users.

DATABASE CONTENTS

To date we have populated the database with >500 samples, which are associated with various diseases (e.g. MS, celiac, HCV, influenza) and tissue types (e.g. brain lesions, lymph nodes, blood) originating from >15 studies. To facilitate standardization of VDJbase data integration, we use a standard pipeline for all genotype and haplotype inferences (see Materials and Methods section for details).

Use case examples

Initial use of this database and annotation system has enabled rigorous testing of human adaptive immune genetic findings. As one example, a number of putative allelic variants reported in older literature appear to be completely absent from Caucasian populations (e.g. IGHJ4*01, IGHJ5*01, IGHJ6*01). Another example is a mosaic deletion pattern that was validated for a much larger cohort (21) (see Figure 2D–F).

Approximately one-third of the population is known to be heterozygous for gene IGHJ6 (21,24), and our database—in which 127 of 326 individuals are heterozygous—is consistent with this observation. We have additionally identified numerous IGH genes to be heterozygous in individuals with a defined genotype: nine IGHV genes observed to be heterozygous in over 50% of individuals (IGHV2-70D, IGHV1-69, IGHV3-53, IGHV3-48, IGHV3-49, IGHV4-30-2, IGHV3-30-3, IGHV3-30 and IGHV3-11); 12 IGHV genes observed to be heterozygous in over 30% of individuals (IGHV3-73, IGHV3-64, IGHV4-61, IGHV1-58, IGHV3-53, IGHV5-51, IGHV4-39, IGHV1-46, IGHV1-18, IGHV3-13, IGHV2-5 and IGHV4-4); and two IGHD genes observed to be heterozygous in 38-65% of individuals (IGHD2-8 and IGHD2-21) (see Figure 2C).

MATERIALS AND METHODS

Inferred genotypes and haplotypes

We align each of the pre-processed datasets using the most recent version of the IgBLAST (32) aligner and the current IMGT germline reference set. We infer previously unknown alleles using TIgGER's inferGenotype function (22) with a modification to the position range input, which allows detection of novel alleles involving sequence variation at nucleotide positions beyond 312 (current default). The sequences are then aligned again, using IgBLAST with a germline reference set that is extended to include any novel IGHV alleles inferred by inferGenotype. The output of IgBLAST is converted to the Change-O format (33) that is compliant with the MiAIRR standard (34). To improve the subsequent quality of allele inference in samples containing highly mutated sequences, we infer clones using SCOPEr (35) and choose a single representative, with the lowest number of mutations, for each clone. A genotype is then inferred using TIgGER's new inferGenotypeBayesian function (26), which can detect novel alleles at greater hamming distance from sequences in the provided reference set than previous versions of TIgGER, and assigns a probability, K_{genotype}, to each allele in the inferred genotype. The sequences are then aligned for a third time with IgBLAST, using a germline reference set that contains only sequences of those alleles included in the personalised genotype created by inferGenotypeBayesian. Lastly, haplotypes are inferred for heterozygous individuals for genes IGHJ6/IGHD2-8/IGHD2-21 using RAbHIT (28). In the current version of TIgGER, up to four distinct alleles are allowed in an individual's genotype. This reflects the possibility of a gene duplication with both loci being heterozygous, as previously observed in this region (17). Samples with fewer than 2000 sequences either in the first IgBLAST or after the collapsing of clones are excluded from further analysis, and data is not incorporated into VDJbase. For datasets with partial V-region coverage, some modifications (described below) are made to the processing protocol. We consistently update the datasets according to the latest versions of the above mentioned tools, and use version control for the website for reproducibility. Versions are displayed on the site.

IGHV annotation for datasets with partial V-region coverage

Aligners are more likely to make ambiguous calls (for example 'IGHV3-23*01 or IGHV3-23*02') when aligning sequences with partial V-region coverage, and this can influence downstream analyses. To resolve this situation, in these datasets, we collapse ambiguous allele assignments using the RAbHIT reliability scores. Each allele for which more than 60% of alignments are ambiguous calls are marked as non-reliable alleles (NRA), and later collapsed. Where ambiguous calls remain, such as 'IGHV3-23*01 or IGHV3-23*02', this is indicated by allele designations such as 01_02. We screen for reliable alleles prior to inferring novel alleles and genotypes. We also modify the numbering of the starting position of partial novel inferences to correspond to the implied nucleotide numbers of full length sequences.

CONCLUSION AND FUTURE DEVELOPMENTS

VDJbase introduces a database for genotypes and haplotypes inferred from AIRR-seq data. These user-friendly web-based queries of VDJbase open new opportunities to browse and extract valuable biological information from the rapidly accumulating AIRR-seq data. In its current form, the focus has been on human B cell receptors, though our database structure is broadly applicable also to T cell receptors, antibody light chains, and Ig and TR from other species. In addition, currently inference of previously unknown alleles is performed by TIgGER. We intend to include options to infer novel alleles using other available tools such as Partis (25) and IgDiscover (23) once there is community agreement on how to reconcile the discrepancies between the results produced by these tools. We hope that the interface between OGRDB and VDJbase will accelerate this process.

VDJbase can shed new light on the variability of germline alleles within and across populations. We anticipate that this database will be extended to effective completeness of human Ig and TR gene variants in all human populations, with enhanced search and analysis features added. Extending the database to include further samples from less sequenced clinical cohorts and ethnic backgrounds should dramatically increase the possibility of discovery of new biomarkers for diseases and treatments.

FUNDING

European Union's Horizon 2020 research and innovation program [825821]. The contents of this document are the sole responsibility of the iReceptor Plus Consortium and can under no circumstances be regarded as reflecting the position of the European Union; ISF [832/16]; NIH [1R01AI127877, 1R01AI125567 and 1R01AI130398]. Funding for open access charge: Horizon 2020 [825821].
Conflict of interest statement. None declared.

REFERENCES

1. 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. *Nat. Biotechnol.*, **32**, 158.
2. Martin,V., Wu,Y.-C., Kipling,D. and Dunn-Walters,D. (2015) Ageing of the B-cell repertoire. *Philos. Trans. R. Soc. B: Biol. Sci.*, **370**, 20140237.
3. Nielsen,S.C., Roskin,K.M., Jackson,K.J., Joshi,S.A., Nejad,P., Lee,J.-Y., Wagar,L.E., Pham,T.D., Hoh,R.A., Nguyen,K.D. *et al.* (2019) Shaping of infant B cell receptor repertoires by environmental factors and infectious disease. *Sci. Transl. Med.*, **11**, eaat2004.
4. Robinson,W.H. (2015) Sequencing the functional antibody repertoire-diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.*, **11**, 171.
5. Stern,J.N., Yaari,G., Vander Heiden,J.A., Church,G., Donahue,W.F., Hintzen,R.Q., Huttner,A.J., Laman,J.D., Nagra,R.M., Nylander,A. *et al.* (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.*, **6**, 248ra107.
6. Snir,O., Mesin,L., Gidoni,M., Lundin,K.E., Yaari,G. and Sollid,L.M. (2015) Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J. Immunol.*, **194**, 5703–5712.
7. Vander Heiden,J.A., Stathopoulos,P., Zhou,J.Q., Chen,L., Gilbert,T.J., Bolen,C.R., Barohn,R.J., Dimachkie,M.M., Ciafaloni,E., Broering,T.J. *et al.* (2017) Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *J. Immunol.*, **198**, 1460–1473.
8. Hoh,R.A., Joshi,S.A., Liu,Y., Wang,C., Roskin,K.M., Lee,J.-Y., Pham,T., Looney,T.J., Jackson,K.J., Dixit,V.P. *et al.* (2016) Single B-cell deconvolution of peanut-specific antibody responses in allergic patients. *J. Allergy Clin. Immunol.*, **137**, 157–167.
9. Tsioris,K., Gupta,N.T., Ogunniyi,A.O., Zimnisky,R.M., Qian,F., Yao,Y., Wang,X., Stern,J.N., Chari,R., Briggs,A.W. *et al.* (2015) Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integr. Biol.*, **7**, 1587–1597.
10. Haynes,B.F., Kelsoe,G., Harrison,S.C. and Kepler,T.B. (2012) B-cell–lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat. Biotechnol.*, **30**, 423.
11. Yaari,G., Vander Heiden,J., Uduman,M., Gadala-Maria,D., Gupta,N., Stern,J.N., O'Connor,K., Hafler,D., Laserson,U., Vigneault,F. *et al.* (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.*, **4**, 358.
12. Yaari,G. and Kleinstein,S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.*, **7**, 121.
13. Avnir,Y., Watson,C.T., Glanville,J., Peterson,E.C., Tallarico,A.S., Bennett,A.S., Qin,K., Fu,Y., Huang,C.-Y., Beigel,J.H. *et al.* (2016) IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Scientific Rep.*, **6**, 20842.
14. Bonsignori,M., Liao,H.-X., Gao,F., Williams,W.B., Alam,S.M., Montefiori,D.C. and Haynes,B.F. (2017) Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. *Immunol. Rev.*, **275**, 145–160.
15. Yacoob,C., Pancera,M., Vigdorovich,V., Oliver,B.G., Glenn,J.A., Feng,J., Sather,D.N., McGuire,A.T. and Stamatatos,L. (2016) Differences in allelic frequency and CDRH3 region limit the engagement of HIV Env immunogens by putative VRC01 neutralizing antibody precursors. *Cell Rep.*, **17**, 1560–1570.
16. Watson,C.T., Matsen,F.A., Jackson,K.J., Bashir,A., Smith,M.L., Glanville,J., Breden,F., Kleinstein,S.H., Collins,A.M. and Busse,C.E. (2017) Comment on 'a database of human immune receptor alleles recovered from population sequencing data'. *J. Immunol.*, **198**, 3371–3373.
17. Watson,C.T., Steinberg,K.M., Huddleston,J., Warren,R.L., Malig,M., Schein,J., Willsey,A.J., Joy,J.B., Scott,J.K., Graves,T.A. *et al.* (2013) Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.*, **92**, 530–546.
18. Watson,C.T., Glanville,J. and Marasco,W.A. (2017) The individual and population genetics of antibody immunity. *Trends Immunol.*, **38**, 459–470.
19. Scheepers,C., Shrestha,R.K., Lambson,B.E., Jackson,K.J., Wright,I.A., Naicker,D., Goosen,M., Berrie,L., Ismail,A., Garrett,N. *et al.* (2015) Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J. Immunol.*, **194**, 4371–4378.
20. Jackson,K. J.L., Wang,Y., Gaeta,B.A., Pomat,W., Siba,P., Rimmer,J., Sewell,W.A. and Collins,A.M. (2012) Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics*, **64**, 3–14.
21. Gidoni,M., Snir,O., Peres,A., Polak,P., Lindeman,I., Mikocziova,I., Sarna,V.K., Lundin,K.E., Clouser,C., Vigneault,F. *et al.* (2019) Mosaic deletion patterns of the human antibody heavy chain gene locus shown by bayesian haplotyping. *Nat. Commun.*, **10**, 628.
22. Gadala-Maria,D., Yaari,G., Uduman,M. and Kleinstein,S.H. (2015) Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E862–E870.
23. Corcoran,M.M., Phad,G.E., Bernat,N.V., Stahl-Hennig,C., Sumida,N., Persson,M.A., Martin,M. and Hedestam,G.B.K. (2016) Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.*, **7**, 13642.
24. Kidd,M.J., Chen,Z., Wang,Y., Jackson,K.J., Zhang,L., Boyd,S.D., Fire,A.Z., Tanaka,M.M., Gaëta,B.A. and Collins,A.M. (2012) The

inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J. Immunol.*, **188**, 1333–1340.
25. Ralph,D.K. and Matsen IV,F.A. (2016) Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.*, **12**, e1004409.
26. Gadala-Maria,D., Gidoni,M., Marquez,S., Vander Heiden,J.A., Kos,J.T., Watson,C.T., OConnor,K., Yaari,G. and Kleinstein,S.H. (2019) Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Front. Immunol.*, **10**, 129.
27. Kirik,U., Greiff,L., Levander,F. and Ohlin,M. (2017) Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol. Immunol.*, **87**, 12–22.
28. Peres,A., Gidoni,M., Polak,P. and Yaari,G. (2019) RAbHIT: R Antibody Haplotype Inference Tool. *Bioinformatics*, doi:10.1093/bioinformatics/btz481.
29. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Jabado-Michaloud,J., Folch,G., Bellahcene,F., Wu,Y., Gemrot,E., Brochet,X., Lane,J. *et al.* (2008) IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Res.*, **37**, D1006–D1012.
30. Ohlin,M., Scheepers,C., Corcoran,M., Lees,W.D., Busse,C.E., Bagnara,D., Thörnqvist,L., Bürckert,J.-P., Jackson,K.J., Ralph,D.K. *et al.* (2019) Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation and naming. *Front. Immunol.*, **10**, 435.
31. Wang,Y., Jackson,K.J., Sewell,W.A. and Collins,A.M. (2008) Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol. Cell Biol.*, **86**, 111–115.
32. Ye,J., Ma,N., Madden,T.L. and Ostell,J.M. (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, **41**, W34–W40.
33. Gupta,N.T., Vander Heiden,J.A., Uduman,M., Gadala-Maria,D., Yaari,G. and Kleinstein,S.H. (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, **31**, 3356–3358.
34. Breden,F., Luning Prak,E.T., Peters,B., Rubelt,F., Schramm,C.A., Busse,C.E., Vander Heiden,J.A., Christley,S., Bukhari,S. A.C., Thorogood,A. *et al.* (2017) Reproducibility and reuse of adaptive immune receptor repertoire data. *Front. Immunol.*, **8**, 1418.
35. Nouri,N. and Kleinstein,S.H. (2018) A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics*, **34**, i341–i349.