Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires

Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M. Deane and Konrad Krawczyk

J Immunol published online 14 September 2018
http://www.jimmunol.org/content/early/2018/09/13/jimmunol.1800708

The Journal of Immunology is published twice each month by The American Association of Immunologists, Inc., 1451 Rockville Pike, Suite 650, Rockville, MD 20852

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

Published September 14, 2018, doi:10.4049/jimmunol.1800708

Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires

Aleksandr Kovaltsuk,* Jinwoo Leem,* Sebastian Kelm,† James Snowden,† Charlotte M. Deane,* and Konrad Krawczyk*

Abs are immune system proteins that recognize noxious molecules for elimination. Their sequence diversity and binding versatility have made Abs the primary class of biopharmaceuticals. Recently, it has become possible to query their immense natural diversity using next-generation sequencing of Ig gene repertoires (Ig-seq). However, Ig-seq outputs are currently fragmented across repositories and tend to be presented as raw nucleotide reads, which means nontrivial effort is required to reuse the data for analysis. To address this issue, we have collected Ig-seq outputs from 55 studies, covering more than half a billion Ab sequences across diverse immune states, organisms (primarily human and mouse), and individuals. We have sorted, cleaned, annotated, translated, and numbered these sequences and make the data available via our Observed Antibody Space (OAS) resource at http://antibodymap.org. The data within OAS will be regularly updated with newly released Ig-seq datasets. We believe OAS will facilitate data mining of immune repertoires for improved understanding of the immune system and development of better biotherapeutics. ***The Journal of Immunology*****, 2018, 201: 000–000.**

Antibodies (or BCR) are protein products of B cells and primary actors of adaptive immunity in-jawed vertebrates (1). They are highly malleable molecules that can bind to virtually any Ag. An organism holds a great variety of these molecules, increasing the probability that an arbitrary Ag can be recognized by an Ab, initiating an immune response (2). Owing to their binding malleability, they are the most prominent class of reagents and biotherapeutics (Refs. 3 and 4 and M. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A.P. Lewis, A. Bujotzek, J. Shi, and C.M. Deane, manuscript posted on bioRxiv). Continued successful exploitation of these molecules relies on our ability to discern the functional diversity of Ab repertoires (5–7).

Next-generation sequencing of Ig gene repertoires (Ig-seq) has enabled researchers to take snapshots of millions of sequences at a time across individuals, diverse organisms, and different immune states (8, 9). The ability to sequence and analyze millions of Ab sequences has the potential to uncover the mechanics of the immune response to any Ag (10, 11) and dysfunctions of the immune system itself (12).

Many previous studies have addressed the issue of Ab diversity, contributing invaluable evidence to understanding the dynamics of human immune systems (13). Numerous analyses have focused on the frequencies of V(D)J gene usages, which can offer insights into creating biased therapeutic Ab libraries (14–16). Another therapeutic application of Ab repertoire analysis is advancing vaccine design by comparative longitudinal studies of pre- and postantigen challenge experiments (10, 11, 17–22). Such comparative studies have shown that different individuals can converge on the same Ab sequence against a given vaccine (11, 19). Because of sequencing limitations, these analyses have focused on H or L chains separately, whereas one ought to study the paired repertoire to obtain deeper insights of Ab diversity (23).

Technical advances in sequencing technology have outpaced storage and analysis pipelines (24, 25). This has meant that the outputs of Ig-seq studies are fragmented across repositories, making it difficult to perform large-scale data mining of Ab repertoires (25). Metadata, such as isotype, age, or subject identifiers, are not typically standardized; therefore, extraction of specific subsets of Ab repertoires for comparative analyses is challenging. Furthermore, the data are typically deposited as raw nucleotide reads. It requires nontrivial ad hoc effort to convert such raw reads to amino acid sequences that ultimately dictate the molecular structure and Ag recognition. Some of these issues are addressed by services that provide Ig-seq–specific data deposition and analysis pipelines such as the B-T.CR wiki (https://b-t.cr), ImmPort (http://immport.org) (26, 27), immunoSEQ Analyzer (http://clients.adaptivebiotech.com/), iReceptor (http://ireceptor.irmacs.sfu.ca/) (28), or VDJServer (http://vdjserver.org) (29). The iReceptor and the VDJServer are the main resources that fall under the umbrella of the organized effort of the Adaptive Immune Receptor Repertoire Community to provide standardized deposition and analysis pipelines for the Ig-seq outputs (24). These services chiefly focus on facilitating bulk deposition of raw data to perform standardized sequencing analyses. Ultimately, because immunoinformatics is not the chief focus of such services, bulk data download from such websites is limited, and converting the raw nucleotide data obtained into a format suitable for analysis still requires installation and running additional software packages. In this study, we identify, clean, annotate, and make the data available as a starting point for immunodiagnostics analyses.

To address these issues, we have created the Observed Antibody Space (OAS) resource that allows large-scale data mining of Ab

*Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom; and †UCB Pharma, Slough SL1 3WE, United Kingdom

ORCIDs: 0000-0002-7817-3644 (J.L.); 0000-0001-7146-9322 (S.K.); 0000-0003-4855-7329 (J.S.); 0000-0003-1388-2252 (C.M.D.).

Received for publication May 22, 2018. Accepted for publication August 19, 2018.

This work was supported by funding from the Biotechnology and Biological Sciences Research Council (Grant BB/M011224/1) and UCB Pharma Ltd., awarded to A.K.

Address correspondence and reprint requests to Prof. Charlotte M. Deane and Dr. Konrad Krawczyk, University of Oxford, Department of Statistics, Oxford OX1 3LB, U.K. E-mail addresses: deane@stats.ox.ac.uk (C.M.D.) and konrad@proteincontact.org (K.K.)

Abbreviations used in this article: CH1, constant heavy domain 1; Ig-seq, next-generation sequencing of Ig gene repertoire; IMGT, International ImMunoGeneTics information system; OAS, Observed Antibody Space.

repertoires. We have, so far, collected the raw outputs of 55 Ig-seq experiments, covering over half a billion sequences. We have organized the sequences by metadata, such as organism, isotype, B cell type, and source, and the immune status of B cell donors to facilitate bulk retrieval of specific subsets for comparative analyses. We have converted all of the Ig-seq sequences to amino acids while preserving the link to the respective original raw nucleotide sequences and numbered them using the International ImMunoGeneTics information system (IMGT) scheme. The data are available for querying or bulk download at http://antibodymap.org. We believe that OAS will facilitate data-mining Ab repertoires for improved understanding of the dynamics of the immune system and, thus, better engineering of biotherapeutics.

Materials and Methods

A list of study accession codes of publicly available Ig-seq datasets were obtained via a literature review. The majority of raw reads were downloaded from the European Nucleotide Archive (30) and the National Center for Biotechnology Information websites (31). In a small number of cases, another public Ig-seq repository was specified [e.g (14, 32–34)]. Metadata were manually extracted from the deposited datasets and arranged in a reproducible format.

The downloaded FASTQ files were processed depending on the sequencing platform. Paired raw Illumina reads were assembled with FLASH (35). The assembled Ab sequences were converted to the FASTA format using FASTX-Toolkit (36). As raw reads from Roche 454 are not paired, these FASTQ files were directly converted to the FASTA format with the FASTX-Toolkit.

The H chain sequences were automatically annotated with isotype information unless such data were given in the corresponding publication. Automatic isotype annotation was performed by aligning the constant heavy domain 1 (CH1) of any given Ab sequence against the IMGT isotype reference (37) of the respective species using the Smith–Waterman (38) algorithm. We assigned a score of 2 for a nucleotide match and a score of −1 for a nucleotide mismatch or a gap. The IMGT isotype references comprised 21-nt–long fragments of the CH1 domain of the Ab isotypes. To ensure a high confidence of correct isotype identification, we employed a

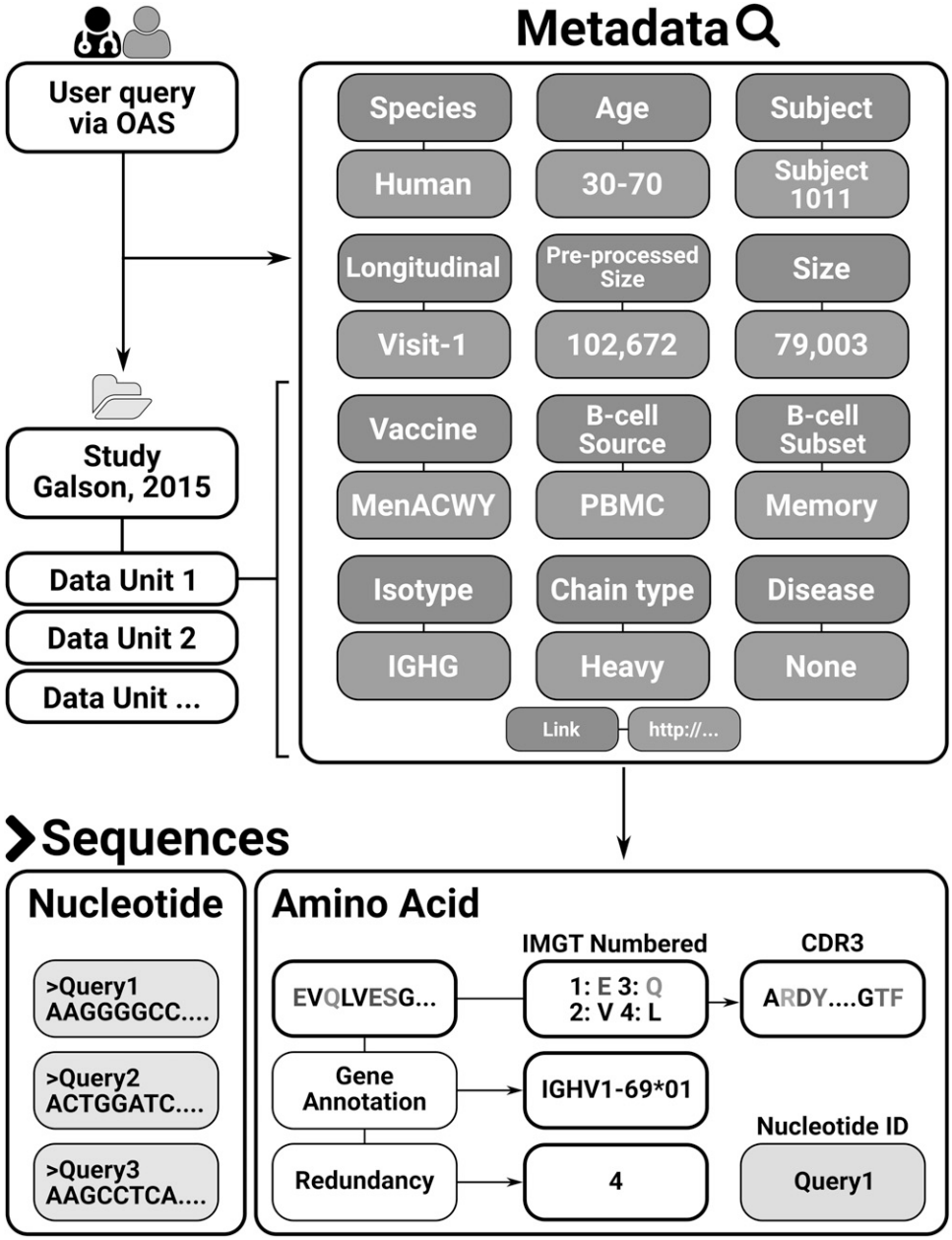

FIGURE 1. The OAS database. The data from 55 studies are sorted into Data Units. Each Data Unit is a set of Ab sequences that share the same set of metadata. Each sequence in a Data Unit is further associated with sequence-specific annotations.

conservative threshold of 30 in the Smith–Waterman algorithm scoring function. Sequences whose Smith–Waterman algorithm score was below the threshold for all isotypes were assigned as "bulk." The robustness of this protocol was confirmed on the author-annotated Ig-seq datasets (18, 39, 40), in which it resulted in 99% accurate annotations. Around 1% of the Ig-seq data had a very short (or missing) CH1 domain sequence. Such sequences were also assigned as bulk.

IgBLASTn (41) was used to convert the FASTA files of Ab nucleotide sequences to amino acids. The amino acid sequences were then parsed with ANARCI (42) using the IMGT scheme (43). In this step, every sequence is IMGT numbered and inspected for compliance with our knowledge of Ig folding. Amino acid sequences that harbor unusual indels in canonical CDR and framework regions or stop codons are removed as these are considered structurally nonviable. ANARCI does not number a sequence if its V and J genes do not align to a Hidden Markov Model (44) built on its respective species amino acid IMGT germlines (37). We also filter out potentially chimeric sequences by detecting duplicated CDR-H3 regions in every amino acid sequence, checking for the complete sequence residue annotation, checking for the full-length framework 4 region, and imposing the length cutoff of 37 residues for CDR-H3 in human, mouse, rat, rabbit, alpaca, and rhesus Abs. Because of technical limitations of sequencing platforms, certain reads were missing significant portions of the V region (e.g., portions of CDR1); sequences that did not have all three CDRs were discarded as incomplete.

The V and J gene annotation available in OAS is obtained using ANARCI, which identifies the germline genes with the highest amino acid identity (37). As V and J genes of camels have not yet been well characterized (45), we employed the alpaca (the closest relative available) Ig genes in camel Ig-seq data interrogation, as these two species belong to the same biological family (Camelidae). If data from other poorly cataloged species are added to OAS, we will use the closest available relative for V and J gene annotation.

Using the protocol above, we annotated Ig-seq results of 55 independent studies. To streamline updating OAS with new data, we have generated a procedure to automatically identify Ig-seq datasets from raw sequence read archives. We apply our Ab annotation protocol to each raw nucleotide dataset deposited in the National Center for Biotechnology Information/European Nucleotide Archive repositories; if we find more than 10,000 Ab sequences in any given dataset, it is set aside for manual inspection. Manual inspection is still necessary to efficiently assign metadata, as these are currently deposited in a nonstandardized manner. This procedure allows for automatic identification of new Ig-seq datasets and semiautomatically updating of OAS.

Results

We collected raw sequencing outputs from 55 Ig-seq studies. All raw nucleotide reads were converted into amino acids using IgBLASTn (41). Within OAS, it is possible to link back from the translated amino acid sequences to the raw nucleotide data. The full amino acid sequences were then IMGT numbered using ANARCI (42). As well as providing IMGT and gene annotations, ANARCI acts as a broad-brush filter of Ab sequences that are likely to be erroneous (see *Materials and Methods*). For each Ig-seq dataset, we provide the total number of amino acids that were retrieved from IgBLASTn outputs as well as after ANARCI parsing. These numbers may be useful as proxies for dataset quality assessment. Applying the same retrieval, amino acid conversion, gene annotation, and numbering protocol to all sequences assures the same point of reference across the 55 heterogeneous Ig-seq datasets (46). This protocol produces the full IMGT-numbered sequences together with gene annotations for each of the 55 datasets.

The numbered amino acid sequences in each dataset are sorted by metadata (e.g., individuals, age, vaccination regimen, B cell type, and source, etc.) (Fig. 1). Deposition of such metadata is currently not standardized and requires ad hoc manual curation for each dataset. In an effort to organize the Ab sequences using such metadata, we have grouped the sequences within each dataset into Data Units. Each Data Unit represents a group of sequences within a given dataset with a unique combination of metadata values. The metadata values are summarized in Table I.

As of July 1, 2018, 55 Ig-seq studies are included in OAS, totaling 618,371,034 sequences (562,544,071 VH and 55,826,963 VL sequences), whereas the total number of translated amino acid sequences that were obtained from IgBLASTn outputs prior to ANARCI parsing is 803,508,673. The majority of the sequences deposited in OAS are murine (~49.4%) and human (~48.4%). Twenty-two of the Ig-seq studies interrogate the immune system of diseased individuals, the most common ailment being HIV (13 studies). The database also contains 24 Ig-seq studies of the naive Ab gene repertoires (the collection of B cells from donors who are healthy and not purposefully vaccinated). The main source of B cells in the OAS database is peripheral blood (~241 mln of sequences), followed by spleen/splenocytes (~198 mln) and bone marrow (~124 mln). The database holds isotype information for each individual heavy sequence, and the two most common isotypes are IgM (~316 mln) and IgG (~144 mln). For ~65 mln sequences, we were not able to assign isotypes with high confidence. The median redundant size of the Ig-seq studies in the OAS database is 2,164,901 sequences, whereas the largest Ig-seq study was that by Greiff et al. (14) (246,449,120 redundant sequences). Two sequences are redundant if they are of identical length and identical amino acid composition. Detailed statistics on each dataset are given in Table II and summary statistics are located at http://antibodymap.org/oasstats. All the data may be bulk downloaded or individual Data Units queried at http://antibodymap.org.

Discussion

In this study, we describe the OAS database, a unified repository to facilitate large-scale data mining of Ab repertoires in both their

Table I. Metadata descriptors of each Data Unit in OAS

Metadata Name	Metadata Description
Chain	H chain/L chain annotation
Isotype	Identified or deposited isotype information
Age	Information on the age of the human B cell donors
Disease	Indication of whether the donor was sick at the time of B cell extraction
Vaccine	Indication if the B cell donor was purposely immunized prior to B cell extraction
B-cell subset	Indication if a particular B cell subset was sorted for Ig-seq
Species	Organism of the B cell donor
B-cell source	Organ/tissue from which the B cells were extracted
Subject	Indication of a particular B cell donor from whom the B cells were sourced
Longitudinal	If the study was longitudinal, an indicator of the time point
Size	Number of redundant amino acid sequences in the Data Unit
Size_igblastn	Number of redundant amino acid sequences extracted from IgBLASTn outputs prior to ANARCI parsing
Link	Link to the source publication

Each Data Unit is uniquely identified by the study and a collection of the metadata values.

Table II. Summary of Ig-seq studies that are incorporated into the OAS database

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Total ANARCI Parsed Sequences
Banerjee et al. (47)	Rabbit	None	HIV	PBMC	Unsorted	4,334,088 (2,926,727)
Bashford-Rogers et al. (48)	Human	CLL/none	None	PBMC	Unsorted	129,013 (86,166)
Bhiman et al. (49)	Human	HIV	None	PBMC	Unsorted	785,751 (187,067)
Bonsignori et al. (50)	Human	HIV/none	None	PBMC	Memory/unsorted	210,377 (57,374)
Collins et al. (51)	Mouse	None	None	Splenocytes	Unsorted	812,439 (194,752)
Corcoran et al. (52)	Human/ mouse/ rhesus	None	None	PBMC	Unsorted	5,307,880 (2,840,877)
Cui et al. (53)	Mouse	None	NP-CGG/none	Splenocytes	Memory	5,513,816 (935,646)
Doria-Rose et al. (13)	Human	HIV	None	PBMC	Unsorted	2,164,901 (549,544)
Ellebedy et al. (54)	Human	None	Flu	PBMC	Naive/memory/ ASC/ABC	9,626,744 (4,807,583)
Fisher et al. (55)	Mouse	None	*Plasmodium*	Spleen	Unsorted	175,015 (113,594)
Galson et al. (18)	Human	None	Hepatitis B	PBMC	Unsorted/plasma cells/Hepatitis B–specific	21,755,739 (10,442,291)
Galson et al. (39)	Human	None	Hepatitis B	PBMC	Unsorted/plasma cells/Hepatitis B–specific	26,687,394 (14,343,236)
Galson et al. (21)	Human	None	Meningitis	PBMC	Naive/plasma cells/ memory/marginal zone	7,918,197 (3,282,907)
Galson et al. (17)	Human	None	Flu	PBMC	Plasma cells	13,685,210 (5,065,786)
Greiff et al. (40)	Mouse	None	NP-CGG	Bone marrow/ spleen	Plasma cells/ plasmablasts	7,955,739 (2,891,649)
Greiff et al. (34)	Mouse	None	NP-CGG	Spleen	ASCs/plasma cells/ naive	788,787 (523,716)
Greiff et al. (14)	Mouse	None	OVA/Hepatitis B/ NP-HEL/none	Spleen/bone marrow	Plasma cells/pre–B cells/naive	246,449,120 (129,417,569)
Gupta et al. (56)	Human	None	Flu/Hepatitis A/ Hepatitis B	PBMC	Unsorted	25,134,322 (9,966,175)
Halliley et al. (57)	Human	None	Flu/tetanus	Bone marrow	Plasma cells	2,348,164 (1,208,616)
Huang et al. (58)	Human	HIV	None	PBMC	Memory	11,693,783 (5,701,433)
Jiang et al. (59)	Human	None	Flu	PBMC	Naive/plasmablasts	3,199,271 (1,809,306)
Joyce et al. (60)	Human	None	None	PBMC	Unsorted	2,747,688 (1,463,421)
Khan et al. (61)	Mouse	None	OVA	Spleen	Unsorted	24,175,033 (7,113,411)
Levin et al. (62)	Human	Allergy	None	PBMC/nasal biopsy	Unsorted	528,173 (370,465)
Levin et al. (63)	Human	Allergy	None	PBMC/bone marrow	Unsorted	29,643,305 (9,557,586)
Li et al. (64)	Camel	None	None	PBMC	Unsorted	1,152,359 (1,127,651)
Liao et al. (65)	Human	HIV	None	PBMC	Unsorted	1,420,314 (619,492)
Lindner et al. (66)	Mouse	None	*Escherichia coli*/ Clostridia/ *Lactobacillus*	Biopsy of small intestine	Unsorted	1,686,350 (544,061)
Meng et al. (67)	Human	CMV/EBV/ none	None	PBMC/lung/spleen/ bone marrow/colon/ jejunum/lymph node/ileum	Unsorted	45,576,606 (21,738,501)
Menzel et al. (68)	Mouse	None	NP-CGG	Spleen/bone marrow	ASCs	14,355,151 (6,058,480)
Mroczek et al. (69)	Human	None	None	PBMC	Immature/ transitional/mature/ plasmacytes/ memory	104,154 (85,525)

(*Table continues*)

Table II. (*Continued*)

Study	Species	Disease	Vaccine	B Cell Source	B Cell Subset	Total ANARCI Parsed Sequences
Ota et al. (70)	Mouse	None	None	Spleen/lymph	Unsorted	21,505 (9,619)
Palanichamy et al. (71)	Human	MS	None	Cerebrospinal fluid/ PBMC	Unsorted	776,895 (292,801)
Parameswaran et al. (11)	Human	Dengue/none/ nondengue febrile illness	None	PBMC	Unsorted	26,584 (23,606)
Prohaska et al. (72)	Mouse	None	None	Spleen/peritoneum	B-1/B-2/marginal zone/follicular	336,723 (198,983)
Rettig et al. (33)	Mouse	None	None	Spleen/splenocytes	Unsorted	41,908 (24,908)
Rubelt et al. (73)	Human	None	None	PBMC	Naive/memory	2,320,947 (1,719,507)
Schanz et al. (32)	Human	HIV/none	None	PBMC	Unsorted	12,734,958 (5,412,549)
Stern et al. (74)	Human	MS	None	Cervical lymph node/white matter lesion/pia mater/ choroid plexus/ cortex/spleen	Unsorted	8,550,247 (3,321,530)
Sundling et al. (75)	Rhesus	None	HIV	PBMC	Unsorted	40,960 (26,298)
Tipton et al. (76)	Human	SLE/none	Flu/tetanus	PBMC	Unsorted	28,204,742 (13,301,396)
Tong et al. (77)	Mouse	None	OVA	Bone marrow/ spleen	Pro–B cells/ follicular	92,936 (56,878)
Turchaninova et al. (78)	Human	None	None	PBMC	Memory/plasma cells/naive	183,949 (176,441)
Vander Heiden et al. (79)	Human	MG/none	None	PBMC	Memory/naive/ unsorted	13,939,166 (5,170,299)
VanDuijn et al. (80)	Rat	None	DNP/HuD	Splenocytes	Unsorted	6,359,396 (4,234,597)
Vergani et al. (81)	Human	None	None	PBMC	Unsorted	14,161,949 (5,987,086)
Wasemann et al. (82)	Mouse	None	NP-CGG	Lamina propria/ bone marrow/ spleen	Unsorted	146,370 (40,132)
Wu et al. (83)	Human	HIV	None	PBMC	Unsorted	394,144 (198,468)
Wu et al. (84)	Human	HIV	None	PBMC	Unsorted	5,545,910 (1,370,109)
Wu et al. (85)	Human	Allergy/none	None	PBMC/nasal biopsy	Unsorted	35,034 (23,923)
Zhou et al. (22)	Human	HIV	None	PBMC	Unsorted	1,541,645 (458,227)
Zhou et al. (86)	Human	HIV	None	PBMC	Unsorted	722,112 (291,670)
Zhu et al. (87)	Human	HIV	None	PBMC	Unsorted	874,930 (174,435)
Zhu et al. (88)	Human	HIV	None	PBMC	Unsorted	1,962,643 (532,350)
Zhu et al. (89)	Human	HIV	None	PBMC	Unsorted	1,290,499 (699,828)

The datasets are organized into studies related to a given Ig-seq experiment. Each study in the OAS database is subdivided into Data Units. Each Data Unit is a collection of IMGT-numbered amino acid sequences uniquely identified by the metadata descriptors given in Table I; five of which (species, disease, vaccine, B cell source, and B cell type) are given in this table. The "Total ANARCI Parsed Sequences" column indicates the total number of redundant sequences in our database, with the nonredundant numbers in parentheses.

ABC, activated B cell; ASC, Ab secreting cell; CLL, chronic lymphocytic leukemia; DNP, keyhole limpet hemocyanine modified with dinitrophenyl; Flu, influenza; HuD, paraneoplastic encephalomyelitis Ag; MG, myasthenia gravis; MS, multiple sclerosis; NP-CGG, chicken γ globulin; NP-HEL, hen egg lysozyme; SLE, systemic lupus erythematosus.

amino acid and nucleotide forms. Absence of well-established repositories in Ig-seq deposition space required us to perform a combination of literature search and manual curation of the datasets to organize the data into OAS. The current lack of widely adopted deposition standards hampers automatic updating of OAS, as datasets in which we find a large number of Abs still require manual curation to perform metadata annotation correctly. Hopefully, efforts such as those by the Adaptive Immune Receptor Repertoire Community will result in standardization of Ig-seq outputs and will further streamline deposition procedures facilitating large-scale data mining of Ab repertoires (24). Devising unified Ab repertoire repositories is challenging because of both the size of the datasets as well as the diverse data descriptors and analytical pipelines desired by bioinformaticians, wet lab scientists, and clinicians (34).

To our knowledge, OAS is the first organized collection of a large body of Ig-seq outputs that is designed for continuous expansion as more and more Ig-seq data become available.The basic data files are stored in an efficiently compressed format and are searchable

by light-weight metadata entries. To allow comparative bioinformatics analyses across different subsets of Ab repertoires, we have annotated the datasets by commonly used metadata descriptions, such as organism, isotype, B cell type and source, and the immune state of B cell donors. To facilitate research about particular Ab sequences or regions, we make full IMGT-numbered, high-quality amino acid sequences available together with gene annotations as well as linked raw nucleotide data.

These data should aid in-depth comparative analyses across different studies to discern the commonalities observed between independent samples as well as directing Ig-seq experiments on not-yet interrogated Ab repertoires. Revealing shared preferences can be invaluable in identifying the portions of the theoretically allowed Ab space that are strategically used to start immune responses (6). Furthermore, such comparative studies can offer a way of deconvoluting the various df of immune repertoires, such as differences between diversity of isotypes (69) or organisms (90). Charting the differences between repertoires of human/mouse is of particular interest for engineering better humanized biotherapeutics (91). Paired Ab chain sequence information provides an enhanced view on Ab biology (92). However, current paired-sequencing approaches only allow for the delineation of CDR-H3 and CDR-L3 sequences (16); as sequencing read length increases to span all three CDR regions, these paired Ig-seq datasets will be incorporated into OAS.

Beyond identifying broad commonalities across repertoires, data mining Ig-seq outputs provides novel avenues for designing better Ab-based therapeutics. The plethora of currently available Ig-seq data offers a glimpse at a set of sequences that should be able to fold and function in an organism. Aligning therapeutic candidates to sequences in Ig-seq repertoires can reveal mutational choices that might be naturally acceptable, hence providing shortcuts for Ab engineering such as humanization (93). Furthermore, contrasting the naturally observed Abs with therapeutic ones can offer insight as to the naturally favored biophysical properties of these molecules (4). All such future applications rely on the availability of well-structured datasets that can offer a unified point of reference for bioinformatics analyses. We hope that OAS will aid data mining Ab repertoires, help identify strategic preferences of our immune systems, and will ultimately improve how we engineer Abs into better therapeutics.

Acknowledgments

We thank all members of Oxford Protein Informatics Group for testing our OAS resource. In particular, we are grateful to Garret M. Morris and Matthew Raybould for comments that significantly improved the quality of our work.

Disclosures

The authors have no financial conflicts of interest.

References

1. Kindt, T. J., R. A. Goldsby, B. A. Osborne, and J. Kuby. 2007. *Kuby Immunology*, 6th Ed., R. A. Goldsby, ed. W.H. Freeman, New York.
2. Glanville, J., W. Zhai, J. Berka, D. Telman, G. Huerta, G. R. Mehta, I. Ni, L. Mei, P. D. Sundar, G. M. R. Day, et al. 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* 106: 20216–20221.
3. Kaplon, H., and J. M. Reichert. 2018. Antibodies to watch in 2018. *MAbs* 10: 183–203.
4. Jain, T., T. Sun, S. Durand, A. Hall, N. R. Houston, J. H. Nett, B. Sharkey, B. Bobrowicz, I. Caffry, Y. Yu, et al. 2017. Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. USA* 114: 944–949.
5. Miho, E., A. Yermanos, C. R. Weber, C. T. Berger, S. T. Reddy, and V. Greiff. 2018. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front. Immunol.* 9: 224.
6. Greiff, V., C. R. Weber, J. Palme, U. Bodenhofer, E. Miho, U. Menzel, and S. T. Reddy. 2017. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immunol.* 199: 2985–2997.
7. Kovaltsuk, A., K. Krawczyk, J. D. Galson, D. F. Kelly, C. M. Deane, and J. Trück. 2017. How B-cell receptor repertoire sequencing can Be enriched with structural antibody data. *Front. Immunol.* 8: 1753.
8. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32: 158–168.
9. Friedensohn, S., T. A. Khan, and S. T. Reddy. 2017. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol.* 35: 203–214.
10. Galson, J. D., A. J. Pollard, J. Trück, and D. F. Kelly. 2014. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol.* 35: 319–331.
11. Parameswaran, P., Y. Liu, K. M. Roskin, K. K. L. Jackson, V. P. Dixit, J. Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, et al. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* 13: 691–700.
12. Ghraichy, M., J. D. Galson, D. F. Kelly, and J. Trück. 2018. B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. *Immunology* 153: 145–160.
13. Doria-Rose, N. A., C. A. Schramm, J. Gorman, P. L. Moore, J. N. Bhiman, B. J. DeKosky, M. J. Ernandes, I. S. Georgiev, H. J. Kim, M. Pancera, et al; NISC Comparative Sequencing Program. 2014. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* 509: 55–62.
14. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T. H. Winkler, and S. T. Reddy. 2017. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.* 19: 1467–1478.
15. Hoi, K. H., and G. C. Ippolito. 2013. Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. *Genes Immun.* 14: 271–276.
16. DeKosky, B. J., T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, and G. Georgiou. 2015. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21: 86–91.
17. Galson, J. D., J. Trück, D. F. Kelly, and R. van der Most. 2016. Investigating the effect of AS03 adjuvant on the plasma cell repertoire following pH1N1 influenza vaccination. *Sci. Rep.* 6: 37229.
18. Galson, J. D., J. Trück, E. A. Clutterbuck, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2016. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. [Published erratum appears in 2016 *Genome Med.* 8: 81.] *Genome Med.* 8: 68.
19. Jackson, K. J. L., Y. Liu, K. M. Roskin, J. Glanville, R. A. Hoh, K. Seo, E. L. Marshall, T. C. Gurley, M. A. Moody, B. F. Haynes, et al. 2014. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* 16: 105–114.
20. Lee, J., D. R. Boutz, V. Chromikova, M. G. Joyce, C. Vollmers, K. Leung, A. P. Horton, B. J. DeKosky, C.-H. Lee, J. J. Lavinder, et al. 2016. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat. Med.* 22: 1456–1464.
21. Galson, J. D., E. A. Clutterbuck, J. Trück, M. N. Ramasamy, M. Münz, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunol. Cell Biol.* 93: 885–895.
22. Zhou, T., J. Zhu, X. Wu, S. Moquin, B. Zhang, P. Acharya, I. S. Georgiev, H. R. Altae-Tran, G. Y. Chuang, M. G. Joyce, et al; NISC Comparative Sequencing Program. 2013. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* 39: 245–258.
23. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, et al. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31: 166–169.
24. Rubelt, F., C. E. Busse, S. A. C. Bukhari, J. P. Bürckert, E. Mariotti-Ferrandiz, L. G. Cowell, C. T. Watson, N. Marthandan, W. J. Faison, U. Hershberg, et al; AIRR Community. 2017. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* 18: 1274–1278.
25. Breden, F., E. T. Luning Prak, B. Peters, F. Rubelt, C. A. Schramm, C. E. Busse, J. A. Vander Heiden, S. Christley, S. A. C. Bukhari, A. Thorogood, et al. 2017. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front. Immunol.* 8: 1418.
26. Bhattacharya, S., S. Andorf, L. Gomes, P. Dunn, H. Schaefer, J. Pontius, P. Berger, V. Desborough, T. Smith, J. Campbell, et al. 2014. ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* 58: 234–239.
27. Bhattacharya, S., P. Dunn, C. G. Thomas, B. Smith, H. Schaefer, J. Chen, Z. Hu, K. A. Zalocusky, R. D. Shankar, S. S. Shen-Orr, et al. 2018. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* 5: 180015.
28. Corrie, B. D., N. Marthandan, B. Zimonja, J. Jaglale, Y. Zhou, E. Barr, N. Knoetze, F. M. W. Breden, S. Christley, J. K. Scott, et al. 2018. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* 284: 24–41.
29. Christley, S., W. Scarborough, E. Salinas, W. H. Rounds, I. T. Toby, J. M. Fonner, M. K. Levin, M. Kim, S. A. Mock, C. Jordan, et al. 2018. VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Front. Immunol.* 9: 976.
30. Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, et al. 2011. The european nucleotide archive. *Nucleic Acids Res.* 39(Database): D28–D31.

31. NCBI Resource Coordinators. 2017. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 45(D1): D12–D17.
32. Schanz, M., T. Liechti, O. Zagordi, E. Miho, S. T. Reddy, H. F. Günthard, A. Trkola, and M. Huber. 2014. High-throughput sequencing of human immunoglobulin variable regions with subtype identification. *PLoS One* 9: e111726.
33. Rettig, T. A., C. Ward, B. A. Bye, M. J. Pecaut, and S. K. Chapes. 2018. Characterization of the naive murine antibody repertoire using unamplified high-throughput sequencing. *PLoS One* 13: e0190982.
34. Greiff, V., P. Bhat, S. C. Cook, U. Menzel, W. Kang, and S. T. Reddy. 2015. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7: 49.
35. Magoč, T., and S. L. Salzberg. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27: 2957–2963.
36. HannonLab. 2014. *FASTX toolkit.* Cold Spring Harbor Laboratory, New York.
37. Giudicelli, V., D. Chaume, and M.-P. Lefranc. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33: D256–D261.
38. Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147: 195–197.
39. Galson, J. D., J. Trück, A. Fowler, E. A. Clutterbuck, M. Münz, V. Cerundolo, C. Reinhard, R. van der Most, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine* 2: 2070–2079.
40. Greiff, V., U. Menzel, U. Haessler, S. C. Cook, S. Friedensohn, T. A. Khan, M. Pogson, I. Hellmann, and S. T. Reddy. 2014. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* 15: 40.
41. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41(W1): W34–W40.
42. Dunbar, J., and C. M. Deane. 2016. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* 32: 298–300.
43. Lefranc, M.-P., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27: 55–77.
44. Eddy, S. R. 1995. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3: 114–120.
45. Arbabi-Ghahroudi, M. 2017. Camelid single-domain antibodies: historical perspective and future outlook. *Front. Immunol.* 8: 1589.
46. Shugay, M., O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, et al. 2014. Towards error-free profiling of immune repertoires. *Nat. Methods* 11: 653–655.
47. Banerjee, S., H. Shi, M. Banasik, H. Moon, W. Lees, Y. Qin, A. Harley, A. Shepherd, and M. W. Cho. 2017. Evaluation of a novel multi-immunogen vaccine strategy for targeting 4E10/10E8 neutralizing epitopes on HIV-1 gp41 membrane proximal external region. *Virology* 505: 113–126.
48. Bashford-Rogers, R. J. M., A. L. Palser, B. J. Huntly, R. Rance, G. S. Vassiliou, G. A. Follows, and P. Kellam. 2013. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* 23: 1874–1884.
49. Bhiman, J. N., C. Anthony, N. A. Doria-Rose, O. Karimanzira, C. A. Schramm, T. Khoza, D. Kitchin, G. Botha, J. Gorman, N. J. Garrett, et al. 2015. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nat. Med.* 21: 1332–1336.
50. Bonsignori, M., T. Zhou, Z. Sheng, L. Chen, F. Gao, M. G. Joyce, G. Ozorowski, G. Y. Chuang, C. A. Schramm, K. Wiehe, et al; NISC Comparative Sequencing Program. 2016. Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. *Cell* 165: 449–463.
51. Collins, A. M., Y. Wang, K. M. Roskin, C. P. Marquis, and K. J. L. Jackson. 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370: 20140236.
52. Corcoran, M. M., G. E. Phad, N. Vázquez Bernat, C. Stahl-Hennig, N. Sumida, M. A. A. Persson, M. Martin, and G. B. Karlsson Hedestam. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* 7: 13642.
53. Cui, A., R. Di Niro, J. A. Vander Heiden, A. W. Briggs, K. Adams, T. Gilbert, K. C. O'Connor, F. Vigneault, M. J. Shlomchik, and S. H. Kleinstein. 2016. A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J. Immunol.* 197: 3566–3574.
54. Ellebedy, A. H., K. J. L. Jackson, H. T. Kissick, H. I. Nakaya, C. W. Davis, K. M. Roskin, A. K. McElroy, C. M. Oshansky, R. Elbein, S. Thomas, et al. 2016. Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat. Immunol.* 17: 1226–1234.
55. Fisher, C. R., H. J. Sutton, J. A. Kaczmarski, H. A. McNamara, B. Clifton, J. Mitchell, Y. Cai, J. N. Dups, N. J. D'Arcy, M. Singh, et al. 2017. T-dependent B cell responses to *Plasmodium* induce antibodies that form a high-avidity multivalent complex with the circumsporozoite protein. *PLoS Pathog.* 13: e1006469.
56. Gupta, N. T., K. D. Adams, A. W. Briggs, S. C. Timberlake, F. Vigneault, and S. H. Kleinstein. 2017. Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *J. Immunol.* 198: 2489–2499.
57. Halliley, J. L., C. M. Tipton, J. Liesveld, A. F. Rosenberg, J. Darce, I. V. Gregoretti, L. Popova, D. Kaminiski, C. F. Fucile, I. Albizua, et al. 2015. Long-lived plasma cells are contained within the CD19(-)CD38(hi)CD138(+) subset in human bone marrow. *Immunity* 43: 132–145.
58. Huang, J., B. H. Kang, E. Ishida, T. Zhou, T. Griesman, Z. Sheng, F. Wu, N. A. Doria-Rose, B. Zhang, K. McKee, et al. 2016. Identification of a CD4-binding-site antibody to HIV that evolved near-*Pan* neutralization breadth. *Immunity* 45: 1108–1121.
59. Jiang, N., J. He, J. A. Weinstein, L. Penland, S. Sasaki, X. S. He, C. L. Dekker, N. Y. Zheng, M. Huang, M. Sullivan, et al. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. [Published erratum appears in 2013 *Sci. Transl. Med.* 5: 193er8.] *Sci. Transl. Med.* 5: 171ra19.
60. Joyce, M. G., A. K. Wheatley, P. V. Thomas, G. Y. Chuang, C. Soto, R. T. Bailer, A. Druz, I. S. Georgiev, R. A. Gillespie, M. Kanekiyo, et al; NISC Comparative Sequencing Program. 2016. Vaccine-induced antibodies that neutralize group 1 and group 2 influenza a viruses. *Cell* 166: 609–623.
61. Khan, T. A., S. Friedensohn, A. R. Gorter de Vries, J. Straszewski, H.-J. Ruscheweyh, and S. T. Reddy. 2016. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* 2: e1501371.
62. Levin, M., J. J. King, J. Glanville, K. J. L. Jackson, T. J. Looney, R. A. Hoh, A. Mari, M. Andersson, L. Greiff, A. Z. Fire, et al. 2016. Persistence and evolution of allergen-specific IgE repertoires during subcutaneous specific immunotherapy. *J. Allergy Clin. Immunol.* 137: 1535–1544.
63. Levin, M., F. Levander, R. Palmason, L. Greiff, and M. Ohlin. 2017. Antibody-encoding repertoires of bone marrow and peripheral blood-a focus on IgE. *J. Allergy Clin. Immunol.* 139: 1026–1030.
64. Li, X., X. Duan, K. Yang, W. Zhang, C. Zhang, L. Fu, Z. Ren, C. Wang, J. Wu, R. Lu, et al. 2016. Comparative analysis of immune repertoires between bactrian Camel's conventional and heavy-chain antibodies. *PLoS One* 11: e0161801.
65. Liao, H. X., R. Lynch, T. Zhou, F. Gao, S. M. Alam, S. D. Boyd, A. Z. Fire, K. M. Roskin, C. A. Schramm, Z. Zhang, et al; NISC Comparative Sequencing Program. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* 496: 469–476.
66. Lindner, C., I. Thomsen, B. Wahl, M. Ugur, M. K. Sethi, M. Friedrichsen, A. Smoczek, S. Ott, U. Baumann, S. Suerbaum, et al. 2015. Diversification of memory B cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nat. Immunol.* 16: 880–888.
67. Meng, W., B. Zhang, G. W. Schwartz, A. M. Rosenfeld, D. Ren, J. J. C. Thome, D. J. Carpenter, N. Matsuoka, H. Lerner, A. L. Friedman, et al. 2017. An atlas of B-cell clonal distribution in the human body. *Nat. Biotechnol.* 35: 879–884.
68. Menzel, U., V. Greiff, T. A. Khan, U. Haessler, I. Hellmann, S. Friedensohn, S. C. Cook, M. Pogson, and S. T. Reddy. 2014. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS One* 9: e96727.
69. Mroczek, E. S., G. C. Ippolito, T. Rogosch, K. H. Hoi, T. A. Hwangpo, M. G. Brand, Y. Zhuang, C. R. Liu, D. A. Schneider, M. Zemlin, et al. 2014. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front. Immunol.* 5: 96.
70. Ota, M., B. H. Duong, A. Torkamani, C. M. Doyle, A. L. Gavin, T. Ota, and D. Nemazee. 2010. Regulation of the B cell receptor repertoire and self-reactivity by BAFF. *J. Immunol.* 185: 4128–4136.
71. Palanichamy, A., L. Apeltsin, T. C. Kuo, M. Sirota, S. Wang, S. J. Pitts, P. D. Sundar, D. Telman, L. Z. Zhao, M. Derstine, et al. 2014. Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Sci. Transl. Med.* 6: 248ra106.
72. Prohaska, T. A., X. Que, C. J. Diehl, S. Hendrikx, M. W. Chang, K. Jepsen, C. K. Glass, C. Benner, and J. L. Witztum. 2018. Massively parallel sequencing of peritoneal and splenic B cell repertoires highlights unique properties of B-1 cell antibodies. *J. Immunol.* 200: 1702–1717.
73. Rubelt, F., C. R. Bolen, H. M. McGuire, J. A. Vander Heiden, D. Gadala-Maria, M. Levin, G. M. Euskirchen, M. R. Mamedov, G. E. Swan, C. L. Dekker, et al. 2016. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* 7: 11112.
74. Stern, J. N. H., G. Yaari, J. A. Vander Heiden, G. Church, W. F. Donahue, R. Q. Hintzen, A. J. Huttner, J. D. Laman, R. M. Nagra, A. Nylander, et al. 2014. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* 6: 248ra107.
75. Sundling, C., Z. Zhang, G. E. Phad, Z. Sheng, Y. Wang, J. R. Mascola, Y. Li, R. T. Wyatt, L. Shapiro, and G. B. Karlsson Hedestam. 2014. Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *J. Immunol.* 192: 3637–3644.
76. Tipton, C. M., C. F. Fucile, J. Darce, A. Chida, T. Ichikawa, I. Gregoretti, S. Schieferl, J. Hom, S. Jenks, R. J. Feldman, et al. 2015. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat. Immunol.* 16: 755–765.
77. Tong, P., A. Granato, T. Zuo, N. Chaudhary, A. Zuiani, S. S. Han, R. Donthula, A. Shrestha, D. Sen, J. M. Magee, et al. 2017. IgH isotype-specific B cell receptor expression influences B cell fate. [Published erratum appears in 2017 *Proc. Natl. Acad. Sci. USA* 114: E9750–E9751.] *Proc. Natl. Acad. Sci. USA* 114: E8411–E8420.
78. Turchaninova, M. A., A. Davydov, O. V. Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I. Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, et al. 2016. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Protoc.* 11: 1599–1616.
79. Vander Heiden, J. A., P. Stathopoulos, J. Q. Zhou, L. Chen, T. J. Gilbert, C. R. Bolen, R. J. Barohn, M. M. Dimachkie, E. Ciafaloni, T. J. Broering, et al. 2017. Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. *J. Immunol.* 198: 1460–1473.
80. VanDuijn, M. M., L. J. Dekker, W. F. J. van IJcken, P. A. E. Sillevis Smitt, and T. M. Luider. 2017. Immune repertoire after immunization as seen by next-generation sequencing and proteomics. *Front. Immunol.* 8: 1286.

81. Vergani, S., I. Korsunsky, A. N. Mazzarello, G. Ferrer, N. Chiorazzi, and D. Bagnara. 2017. Novel method for high-throughput full-length IGHV-D-J sequencing of the immune repertoire from bulk B-cells with single-cell resolution. *Front. Immunol.* 8: 1157.
82. Wesemann, D. R., A. J. Portuguese, R. M. Meyers, M. P. Gallagher, K. Cluff-Jones, J. M. Magee, R. A. Panchakshari, S. J. Rodig, T. B. Kepler, and F. W. Alt. 2013. Microbial colonization influences early B-lineage development in the gut lamina propria. *Nature* 501: 112–115.
83. Wu, X., T. Zhou, J. Zhu, B. Zhang, I. Georgiev, C. Wang, X. Chen, N. S. Longo, M. Louder, K. McKee, et al; NISC Comparative Sequencing Program. 2011. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333: 1593–1602.
84. Wu, X., Z. Zhang, C. A. Schramm, M. G. Joyce, Y. D. Kwon, T. Zhou, Z. Sheng, B. Zhang, S. O'Dell, K. McKee, et al; NISC Comparative Sequencing Program. 2015. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* 161: 470–485.
85. Wu, Y. C. B., L. K. James, J. A. Vander Heiden, M. Uduman, S. R. Durham, S. H. Kleinstein, D. Kipling, and H. J. Gould. 2014. Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. *J. Allergy Clin. Immunol.* 134: 604–612.
86. Zhou, T., R. M. Lynch, L. Chen, P. Acharya, X. Wu, N. A. Doria-Rose, M. G. Joyce, D. Lingwood, C. Soto, R. T. Bailer, et al; NISC Comparative Sequencing Program. 2015. Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell* 161: 1280–1292.
87. Zhu, J., S. O'Dell, G. Ofek, M. Pancera, X. Wu, B. Zhang, Z. Zhang, J. C. Mullikin, M. Simek, D. R. Burton, et al; NISC Comparative Sequencing Program. 2012. Somatic populations of PGT135–137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatic. *Front. Microbiol.* 3: 315.
88. Zhu, J., G. Ofek, Y. Yang, B. Zhang, M. K. Louder, G. Lu, K. McKee, M. Pancera, J. Skinner, Z. Zhang, et al; NISC Comparative Sequencing Program. 2013. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. USA* 110: 6470–6475.
89. Zhu, J., X. Wu, B. Zhang, K. McKee, S. O'Dell, C. Soto, T. Zhou, J. P. Casazza, J. C. Mullikin, P. D. Kwong, et al; NISC Comparative Sequencing Program. 2013. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc. Natl. Acad. Sci. USA* 110: E4088–E4097.
90. Schroeder, H. W., Jr. 2006. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* 30: 119–135.
91. Zemlin, M., M. Klinger, J. Link, C. Zemlin, K. Bauer, J. A. Engler, H. W. Schroeder, Jr., and P. M. Kirkham. 2003. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol. Biol.* 334: 733–749.
92. DeKosky, B. J., O. I. Lungu, D. Park, E. L. Johnson, W. Charab, C. Chrysostomou, D. Kuroda, A. D. Ellington, G. C. Ippolito, J. J. Gray, and G. Georgiou. 2016. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl. Acad. Sci. USA* 113: E2636–E2645.
93. Olimpieri, P. P., P. Marcatili, and A. Tramontano. 2015. Tabhu: tools for antibody humanization. *Bioinformatics* 31: 434–435.