Title: Observed Antibody Space: a resource for data mining next generation sequencing antibody repertoires.

Aleksandr Kovaltsuk[1], Jinwoo Leem[1], Sebastian Kelm[2], James Snowden[2], Charlotte M. Deane[1,*], Konrad Krawczyk[1,*]

1. University of Oxford, Department of Statistics, Oxford, UK
2. UCB Pharma, Slough, UK

* to whom the correspondence should be addressed,
deane@stats.ox.ac.uk
konrad@proteincontact.org

Abstract. Antibodies are immune system proteins that recognize noxious molecules for elimination. Their sequence diversity and binding versatility have made antibodies the primary class of biopharmaceuticals. Recently it has become possible to query their immense natural diversity using next-generation sequencing of immunoglobulin 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 53 studies, covering more than half a billion antibody sequences across diverse immune states, organisms 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 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.

1. Introduction

Antibodies (or B-cell receptors) 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 antigen. An organism holds a great variety of these molecules increasing the probability that an arbitrary antigen can be recognized by an antibody, initiating an immune response [2]. Owing to their binding malleability

they are the most prominent class of reagents and biotherapeutics [3,4]. Continued successful exploitation of these molecules relies on our ability to discern the functional diversity of antibody repertoires [5–7].

Next-generation sequencing of immunoglobulin 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 analyze millions of antibody sequences has the potential to uncover the mechanics of the immune response to any antigen [10,11] and dysfunctions of the immune system itself [12].

Many previous studies have addressed the issue of antibody diversity, contributing invaluable evidence to 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 antibody libraries for therapeutic applications [14–16]. Another therapeutic application of antibody repertoire analysis is advancing vaccine design by comparative longitudinal studies of pre- and post-antigen challenge experiments [10,11,17–22]. Such comparative studies have shown that different individuals can converge on the same antibody sequence against a given vaccine [11,19]. Due to sequencing limitations, these analyses have focused on heavy or light chains separately, whereas one ought to study the paired repertoire to obtain a deeper insight of the antibody 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 antibody repertoires [25]. Metadata such as isotype, age or subject identifiers are not typically standardized, therefore extraction of specific subsets of antibody repertoires for comparative analyses is challenging. Furthermore, the data are typically deposited as raw nucleotide reads. It requires non-trivial *ad hoc* effort to convert such raw reads to amino acid sequences that ultimately dictate the molecular structure and antigen-recognition. Some of these issues are addressed by services that provide Ig-seq-specific data deposition and analysis pipelines such as Immport (http://immport.org) [26,27], or ImmunoseqAnalyzer (http://clients.adaptivebiotech.com/), IReceptor (http://ireceptor.irmacs.sfu.ca/) or VDJServer (http://vdjserver.org) [28]. The IReceptor and the VDJServer are the main resources that fall under the umbrella of the organized effort by the Adaptive Immune Receptor Repertoire (AIRR) 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

64 services, bulk data download from such websites is limited and converting the raw nucleotide data
65 obtained into format suitable for analysis still requires non-trivial effort.
66
67 To address these issues, we have created the Observed Antibody Space (OAS) resource that allows large-
68 scale data mining of antibody repertoires. We have collected the raw outputs of 53 Ig-seq experiments
69 covering over half a billion sequences. We have organized the sequences by metadata such as organism,
70 isotype, B-cell type and source, and the immune status of B-cell donors to facilitate bulk retrieval of
71 specific subsets for comparative analyses. We have converted all of the Ig-seq sequences to amino acids
72 and numbered them using the IMGT scheme. The data is available for querying or bulk download at
73 http://antibodymap.org. We believe that OAS will facilitate data mining antibody repertoires for
74 improved understanding of the dynamics of the immune system and thus better engineering of
75 biotherapeutics.
76

77 **2. Materials and Methods**
78

79 A list of study accession codes of publically available Ig-seq datasets were obtained via a literature
80 review. The majority of raw reads were downloaded from the European Nucleotide Archive (ENA) [29] and
81 the National Center for Biotechnology Information (NCBI) websites [30]. In a small number of cases,
82 another public Ig-seq repository was specified e.g. [14,31–33]. Metadata were manually extracted from the
83 deposited datasets and arranged in a reproducible format.
84
85 The downloaded FASTQ files were processed depending on the sequencing platform. Paired raw
86 Illumina reads were assembled with FLASH [34]. The assembled antibody sequences were converted to the
87 FASTA format using FASTX-toolkit [35]. As raw reads from Roche 454 are not paired, these FASTQ files
88 were directly converted to the FASTA format with the FASTX-toolkit.
89
90 The heavy chain sequences were automatically annotated with isotype information unless such data was
91 given in the corresponding publication. Automatic isotype annotation was performed by aligning the
92 constant heavy domain 1 (CH1) of any given antibody sequence against the IMGT isotype reference [36] of
93 the respective species using the Smith-Waterman algorithm [37]. We assigned score of two for a
94 nucleotide match, and a score of minus one for a nucleotide mismatch or a gap in our Smith-Waterman
95 scoring function. The IMGT isotype references comprised 21 nucleotide-long fragments of the CH1

domain of the antibody isotypes. To ensure a high confidence of correct isotype identification, we employed a 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 [38–40] where it resulted in 99% accurate annotations. Around 1% of the Ig-seq data had a very short (or missing) sequence of CH1 domain. Such sequences were also assigned as 'bulk'.

IgBlastn [41] was used to convert the FASTA files of antibody nucleotide sequences to amino acids. The amino acid sequences were then numbered with ANARCI [42] using the IMGT scheme [43]. ANARCI does not number a sequence if it does not align to a suitable Hidden Markov Model [44]. ANARCI therefore ensures that antibody sequences do not harbor unusual indels or stop codons in the antibody regions, that V and J genes align to respective species amino acid IMGT germlines [36], and that the length of CDR-H3 is not greater than 37 residues in human, mouse, rat, rabbit, alpaca and rhesus antibodies. Due to technical limitations of sequencing platforms, certain reads were missing significant portions of the variable region (e.g. portions of CDR1), sequences that did not have all three CDRs were discarded as incomplete. The V and J genes are identified during the ANARCI numbering step.

Using the protocol above we annotated Ig-seq results of 53 independent studies. In order 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 antibody annotation protocol to each raw nucleotide dataset deposited in the NCBI/ENA repositories, if we find more than 10,000 antibody sequences in any given dataset, it is set aside for manual inspection. It is still necessary to manually inspect raw datasets to efficiently assign metadata as these are currently deposited in a non-standardized manner. This procedure allows for automatic identification of new Ig-seq datasets, which will enable us to semi automatically update OAS.

3. Results

We collected raw sequencing outputs from 53 Ig-seq studies. All raw nucleotide reads were converted into amino acids using IgBlastn [41]. 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 antibody sequences that are likely to be erroneous (see Materials and Methods). Applying the same

retrieval, amino acid conversion, gene annotation and numbering protocol to all sequences assures the same point of reference across the 53 heterogeneous Ig-seq datasets [45]. This protocol produces the full IMGT-numbered sequences together with gene annotations for each of the 53 datasets.

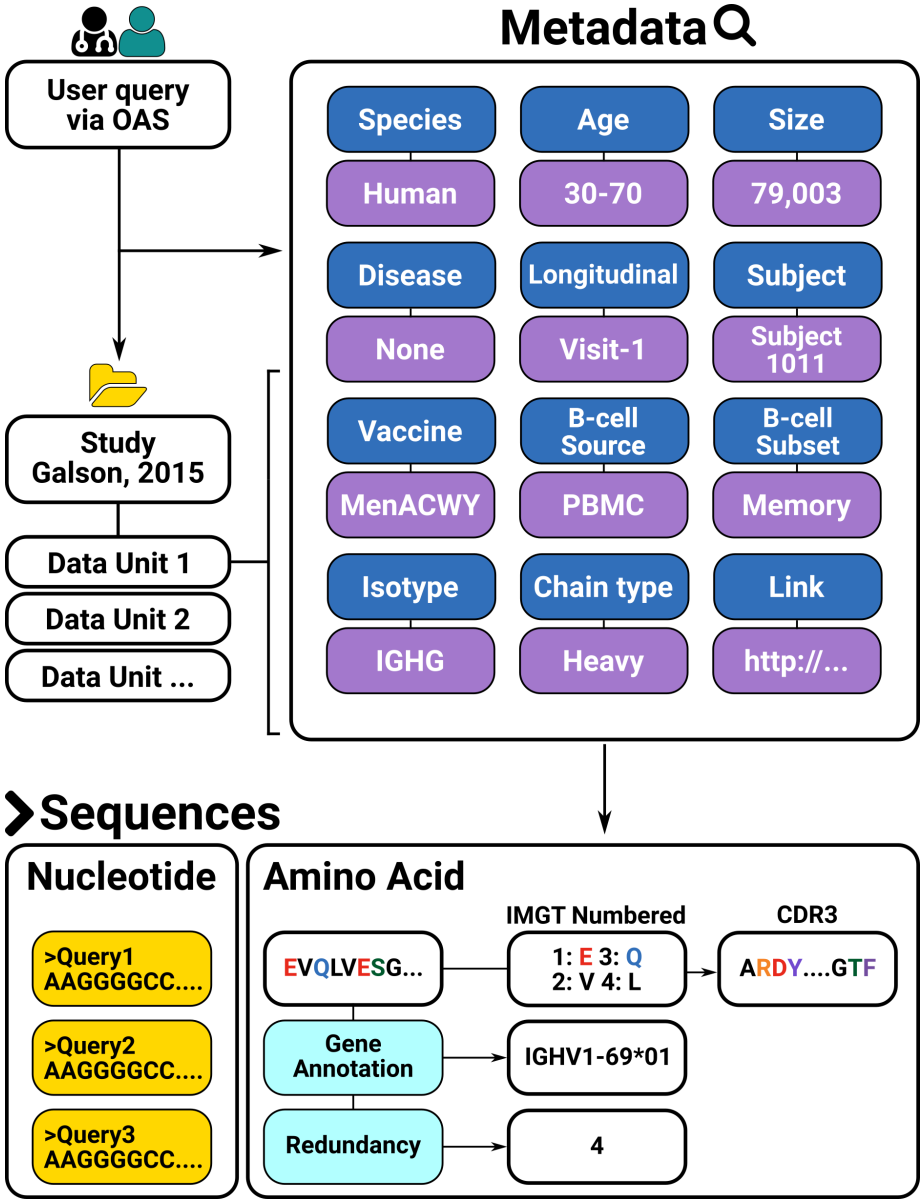

Figure 1. T**he Observed Antibody Space database**. The data from 53 studies is sorted into Data Units. Each Data Unit is a set of antibody sequences that share the same set of meta-data. Each sequence in a Data Unit is further associated with sequence-specific annotations.

138 The numbered amino acid sequences in each dataset are sorted by metadata e.g. individuals, age,
139 vaccination regime, B-cell type and source *etc*. (Figure 1). Deposition of such metadata is currently not
140 standardized and requires *ad hoc* manual curation for each dataset. In an effort to organize the antibody
141 sequences using such metadata, we have grouped the sequences within each dataset into Data Units.
142 Each Data Unit represents a group of sequences within a given dataset with a unique combination of
143 metadata values. The metadata values are summarized in Table 1.
144

Metadata name	**Metadata description**
Chain	Heavy/light 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	Which organ/tissue the B-cells were extracted from.
Subject	Indication of a particular B-cell donor the B-cells were sourced from.
Longitudinal	If the study was longitudinal, an indicator of the time point.
Size	Number of non-redundant sequences in the dataset.
Link	Link to the source publication.

145 Table 1. Metadata descriptors of each Data Unit in OAS. Each Data Unit is uniquely identified by the
146 study and a collection of the metadata values.
147
148
149
150
151
152
153
154

155 As of April 29th 2018, 53 Ig-seq studies are included in OAS totaling 608,651,423 sequences (552,824,460
156 VH and 55,826,963 VL sequences). The majority of these sequences are murine (~50.4%) and human
157 (~47.4%). Twenty-two of the Ig-seq studies interrogate the immune system of diseased individuals, the
158 most common ailment being HIV (13 studies). The database also contains 22 Ig-seq studies of the naive
159 antibody gene repertoires (the collection of B-cells from donors who are healthy and not purposefully
160 vaccinated). The main source of B-cells in the OAS database is peripheral blood (~231m of sequences)
161 followed by spleen/splenocytes (~198m) and bone marrow (~124m). The database holds isotype
162 information for each individual heavy sequence and the two most common isotypes are IgM (~312m)
163 and IgG (~139m). For ~65m sequences we were not able to assign isotypes with high confidence. The
164 median total redundant size of the Ig-seq studies in the OSA database is 2,006,196 sequences, while the
165 largest Ig-seq study was that by Greiff et al., (246,449,189 redundant sequences) [14]. Detailed statistics on
166 each dataset are given in Table 2. All the data may be bulk downloaded or individual Data Units queried,
167 at http://antibodymap.org.
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185

Study	Species	Disease	Vaccine	B-cell source	B-cell subset	Total ANARCI parsed sequences
Banerjee et al., [46]	Rabbit	None	HIV	PBMC	Unsorted	4,334,088 (2,926,727)
Bashford-Rogers et al., [47]	Human	CLL/None	None	PBMC	Unsorted	129,013 (86,166)
Bhiman et al., [48]	Human	HIV	None	PBMC	Unsorted	785,751 (187,067)
Bonsignori et al., [49]	Human	HIV/None	None	PBMC	Memory/ Unsorted	210,377 (57,374)
Collins et al., [50]	Mouse	None	None	Splenocytes	Unsorted	812,439 (194,752)
Corcoran et al., [51]	Human/ Mouse/ Rhesus	None	None	PBMC	Unsorted	5,307,880 (2,840,877)
Cui et al., [52]	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)
Fisher et al., [53]	Mouse	None	Plasmodium	Spleen	Unsorted	175,015 (113,594)
Galson et al., [38]	Human	None	Hepatitis-B	PBMC	Unsorted/ Plasma cells/ HepB-specific	21,755,739 (10,442,291)
Galson et al., [39]	Human	None	Hepatitis-B	PBMC	Unsorted/ Plasma cells/ HepB-specific	26,687,394 (14,343,236)
Galson et al., [21]	Human	None	Meningitis	PBMC	Naïve/ 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., [33]	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	246,449,189 (129,417,638)
Gupta et al., [54]	Human	None	Flu/Hepatitis-A/ Hepatitis-B	PBMC	Unsorted	25,134,322 (9,966,175)
Halliley et al., [55]	Human	None	Flu/Tetanus	Bone marrow	Plasma cells	2,348,164 (1,208,616)
Huang et al., [56]	Human	HIV	None	PBMC	Memory	11,693,783 (5,701,433)
Jiang et al., [57]	Human	None	Flu	PBMC	Naïve/ Plasmablasts	3,199,271 (1,809,306)
Joyce et al., [58]	Human	None	None	PBMC	Unsorted	2,747,688 (1,463,421)
Khan et al., [59]	Mouse	None	OVA	Spleen	Unsorted	24,175,033 (7,113,411)
Levin et al., [60]	Human	Allergy	None	PBMC/ Nasal biopsy	Unsorted	528,173 (370,465)
Levin et al., [61]	Human	Allergy	None	PBMC/	Unsorted	29,643,305

				Bone marrow		(9,557,586)
Li et al., [62]	Camel	None	None	PBMC	Unsorted	1,152,359 (1,127,651)
Liao et al., [63]	Human	HIV	None	PBMC	Unsorted	1,420,314 (619,492)
Lindner et al., [64]	Mouse	None	E.Coli/ Clostridia/ Lactobacillus	Biopsy of small intestine	Unsorted	1,686,350 (544,061)
Meng et al., [65]	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., [66]	Mouse	None	NP-CGG	Spleen/Bone marrow	ASCs	14,355,151 (6,058,480)
Mroczek et al., [67]	Human	None	None	PBMC	Immature/ Transitional/ Mature/ Plasmacytes/ Memory	104,154 (85,525)
Ota et al., [68]	Mouse	None	None	Spleen/Lymph	Unsorted	21,505 (9,619)
Palanichamy et al., [69]	Human	MS	None	CSF/PBMC	Unsorted	776,895 (292,801)
Parameswaran et al., [11]	Human	Dengue/None / Non-dengue febrile illness/	None	PBMC	Unsorted	26,584 (23,606)
Prohaska et al., [70]	Mouse	None	None	Spleen/ Peritoneum	B-1/ B-2/ Marginal Zone/ Follicular	336,723 (198,983)
Rettig et al., [32]	Mouse	None	None	Spleen/ Splenocytes	Unsorted	41,908 (24,908)
Rubelt et al., [71]	Human	None	None	PBMC	Naïve/Memory	2,320,947 (1,719,507)
Schanz et al., [31]	Human	HIV/None	None	PBMC	Unsorted	12,734,958 (5,412,549)
Zhu et al., [72]	Human	HIV	None	PBMC	Unsorted	1,962,643 (532,350)
Stern et al., [73]	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., [74]	Rhesus	None	HIV	PBMC	Unsorted	40,960 (26,298)
Tipton et al., [75]	Human	SLE/None	Flu/Tetanus	PBMC	Unsorted	28,204,742 (13,301,396)
Turchaninova et al., [76]	Human	None	None	PBMC	Memory/Plasma cells/Naive	183,949 (176,441)
Vander Heiden et al., [77]	Human	MG/None	None	PBMC	Memory/Naïve/ Unsorted	13,939,166 (5,170,299)
VanDuijn et al., [78]	Rat	None	DNP/HuD	Splenocytes	Unsorted	6,359,396 (4,234,597)
Vergani et al., [79]	Human	None	None	PBMC	Unsorted	14,161,949 (5,987,086)

Wasemann et al., [80]	Mouse	None	NP-CGG	Lamina propria/ Bone marrow/ Spleen	Unsorted	146,370 (40,132)
Wu et al., [81]	Human	HIV	None	PBMC	Unsorted	394,144 (198,468)
Wu et al., [82]	Human	HIV	None	PBMC	Unsorted	5,545,910 (1,370,109)
Wu et al., [83]	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., [84]	Human	HIV	None	PBMC	Unsorted	722,112 (291,670)
Zhu et al., [85]	Human	HIV	None	PBMC	Unsorted	874,930 (174,435)
Zhu et al., [86]	Human	HIV	None	PBMC	Unsorted	1,290,499 (699,828)

186 Table 2. **Summary of Ig-seq studies that are incorporated into the Observed Antibody Space database.**
187 We organized the datasets among studies related to a given Ig-seq experiment. Each study in the OAS
188 database is subdivided into Data Units. Each Data Unit is a collection of IMGT-numbered amino acid
189 sequences uniquely identified by the metadata descriptors given in Table 1, five of which (species,
190 disease, vaccine, B-cell source and B-cell type) are given in this Table. The 'total ANARCI parsed
191 sequences' field indicates the total number of redundant sequences in our database, with the non-
192 redundant numbers in parentheses. Abbreviations: PBMC, peripheral blood mononuclear cell; CLL,
193 chronic lymphocytic leukemia; NP-CGG, chicken gamma globulin; ASC, antibody secreting cell; OVA,
194 ovalbumin; NP-HEL, hen egg lysozyme; CSF, cerebrospinal fluid; MS, multiple sclerosis; SLE, systemic
195 lupus erythematosus; MG, myasthenia gravis; DNP, dinitrophenyl; HuD, paraneoplastic
196 encephalomyelitis antigen.
197
198
199
200
201
202
203
204
205
206
207
208

209 **4. Discussion**
210
211 Here, we describe the Observed Antibody Space (OAS) database, a unified repository to facilitate large-
212 scale data mining of antibody repertoires in both their amino acid and nucleotide forms. Absence of well
213 established repositories in Ig-seq deposition space required us to perform a combination of literature
214 search and manual curation of the datasets in order to organize the data into OAS. The current lack of
215 widely-adopted deposition standards hampers automatic updating of OAS, as datasets where we find
216 large number of antibodies still require manual curation to perform metadata annotation correctly.
217 Hopefully, efforts such as that by the AIRR community will result in standardization of Ig-seq outputs
218 and will further streamline deposition procedures facilitating large-scale data mining of antibody
219 repertoires [24]. Devising unified antibody repertoire repositories is challenging due to both the size of the
220 datasets as well as the diverse data descriptors and analytical pipelines desired by bioinformaticians,
221 wetlab scientists and clinicians [33].
222
223 OAS is the first organized collection of a large body of Ig-seq outputs. In order to allow comparative
224 bioinformatics analyses across different subsets of the antibody repertoires, we have annotated the
225 datasets by commonly used metadata descriptions such as organism, isotype, B-cell type and source,
226 and the immune state of B-cell donors. To facilitate research about particular antibody sequences or
227 regions, we make full IMGT-numbered high-quality amino acid sequences available together with gene
228 annotations, as well as raw nucleotide data.
229
230 This data should aid in-depth comparative analyses across different studies to discern the commonalities
231 observed between independent samples as well as directing Ig-seq experiments on not yet interrogated
232 antibody repertoires. Revealing shared preferences can be invaluable in identifying the portions of the
233 theoretically allowed antibody space that are strategically used to start an immune response [6].
234 Furthermore, such comparative studies can offer a way of deconvoluting the various degrees of freedom
235 of immune repertoires such as differences between diversity of isotypes [67] or organisms [87]. Charting the
236 differences between repertoires of human/mouse is of particular interest for engineering better
237 humanized biotherapeutics [88].
238
239 Beyond identifying broad commonalities across repertoires, data mining Ig-seq outputs provides novel
240 avenues for designing better antibody-based therapeutics. The plethora of currently available Ig-seq

241 data offers a glimpse at a set of sequences that should be able to fold and function in an organism.
242 Aligning therapeutic candidates to sequences in Ig-seq repertoires can reveal mutational choices that
243 might be naturally acceptable hence providing shortcuts for antibody engineering such as humanization
244 [89]. Furthermore, contrasting the naturally observed antibodies with therapeutic ones can offer insight as
245 to the naturally favored biophysical properties of these molecules [4,90,91]. All such future applications rely
246 on the availability of well-structured datasets that can offer a unified point of reference for
247 bioinformatics analyses. We hope that OAS will aid data mining antibody repertoires, help identify
248 strategic preferences of our immune systems and will ultimately improve how we engineer antibodies
249 into better therapeutics.
250

251 **Acknowledgment:**
252 We would like to thank all members of Oxford Protein Informatics Group for testing our OAS resource.
253 In particular, we are grateful to Garret M. Morris and Matthew Raybould for their comments, which
254 significantly improved the quality of our work.
255

256 **Funding:**
257 This work was supported by funding from Biotechnology and Biological Sciences Research Council
258 (BBSRC) [BB/M011224/1] and UCB Pharma Ltd awarded to AK.

259 **References:**
260
261 1. Kindt, T. J., Goldsby, R. A., Osborne, B. A. & Kuby, J. *Kuby immunology*. (WH Freeman & Company,
262 2007).
263 2. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives
264 insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 20216–
265 20221 (2009).
266 3. Kaplon, H. & Reichert, J. M. Antibodies to watch in 2018. *mAbs* 1–21 (2018).
267 doi:10.1080/19420862.2018.1415671
268 4. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.*
269 **114,** (2017).
270 5. Miho, E. *et al.* Computational Strategies for Dissecting the High-Dimensional Complexity of
271 Adaptive Immune Repertoires. *Front. Immunol.* **9,** (2018).

272 6. Greiff, V. *et al.* Learning the High-Dimensional Immunogenomic Features That Predict Public and
273 Private Antibody Repertoires. *J. Immunol.* ji1700594 (2017). doi:10.4049/jimmunol.1700594
274 7. Kovaltsuk, A. *et al.* How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural
275 Antibody Data. *Front. Immunol.* **8,** 1753 (2017).
276 8. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody
277 repertoire. *Nat. Biotechnol.* **32,** 158–68 (2014).
278 9. Friedensohn, S., Khan, T. A. & Reddy, S. T. Advanced Methodologies in High-Throughput
279 Sequencing of Immune Repertoires. *Trends in Biotechnology* **35,** 203–214 (2017).
280 10. Galson, J., Pollard, A. J., Trück, J. & Kelly, D. F. Studying the antibody repertoire after vaccination:
281 Practical applications. *Trends in Immunology* **35,** 319–331 (2014).
282 11. Parameswaran, P. *et al.* Convergent antibody signatures in human dengue. *Cell Host Microbe* **13,**
283 691–700 (2013).
284 12. Ghraichy, M., Galson, J. D., Kelly, D. F. & Trück, J. B-cell receptor repertoire sequencing in patients
285 with primary immunodeficiency: A review. *Immunology* 145–160 (2017). doi:10.1111/imm.12865
286 13. Doria-Rose, N. A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing
287 antibodies. *Nature* **508,** 55–62 (2014).
288 14. Greiff, V. *et al.* Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of
289 Antibody Repertoires throughout B Cell Development. *Cell Rep.* **19,** 1467–1478 (2017).
290 15. Hoi, K. H. & Ippolito, G. C. Intrinsic bias and public rearrangements in the human immunoglobulin
291 Vλ light chain repertoire. *Genes Immun.* 1–6 (2013). doi:10.1038/gene.2013.10
292 16. Dekosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy- and light-
293 chain antibody repertoire. *Nat. Med.* **21,** 1–8 (2014).
294 17. Galson, J., Trück, J., Kelly, D. F. & van der Most, R. Investigating the effect of AS03 adjuvant on
295 the plasma cell repertoire following pH1N1 influenza vaccination. *Sci. Rep.* **6,** 37229 (2016).
296 18. Galson, J. *et al.* B cell repertoire dynamics after sequential Hepatitis B vaccination, and evidence
297 for cross-reactive B cell activation. *Submitt. Manuscr.* 1–13 (2016). doi:10.1186/s13073-016-
298 0322-z
299 19. Jackson, K. J. L. *et al.* Human responses to influenza vaccination show seroconversion signatures
300 and convergent antibody rearrangements. *Cell Host Microbe* **16,** 105–114 (2014).
301 20. Lee, J. *et al.* Molecular-level analysis of the serum antibody repertoire in young adults before and
302 after seasonal influenza vaccination. *Nat Med* **22,** 1456–1464 (2016).
303 21. Galson, J. *et al.* BCR repertoire sequencing: Different patterns of B-cell activation after two

304 Meningococcal vaccines. *Immunol. Cell Biol.* **93,** 885–895 (2015).
305 22. Zhou, T. *et al.* Multidonor analysis reveals structural elements, genetic determinants, and
306 maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39,** 245–258
307 (2013).
308 23. DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and
309 light chain repertoire. *Nat. Biotechnol.* **31,** 166–9 (2013).
310 24. Rubelt, F. *et al.* Adaptive Immune Receptor Repertoire Community recommendations for sharing
311 immune-repertoire sequencing data. *Nature Immunology* **18,** 1274–1278 (2017).
312 25. Breden, F. *et al.* Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front.*
313 *Immunol.* **8,** 1418 (2017).
314 26. Bhattacharya, S. *et al.* ImmPort: Disseminating data to the public for the future of immunology.
315 *Immunol. Res.* **58,** 234–239 (2014).
316 27. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for
317 translational and clinical research. *Sci. Data* **5,** (2018).
318 28. Cowell, L. G. VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune
319 Repertoire Sequences and Rearrangements. *Front. Immunol.* **9,** 976 (2018).
320 29. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* **39,** (2011).
321 30. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology
322 Information. *Nucleic Acids Res.* **45,** D12–D17 (2017).
323 31. Schanz, M. *et al.* High-throughput sequencing of human immunoglobulin variable regions with
324 subtype identification. *PLoS One* **9,** (2014).
325 32. Rettig, T. A., Ward, C., Bye, B. A., Pecaut, M. J. & Chapes, S. K. Characterization of the naive
326 murine antibody repertoire using unamplified high-throughput sequencing. *PLoS One* **13,** (2018).
327 33. Greiff, V. *et al.* A bioinformatic framework for immune repertoire diversity profiling enables
328 detection of immunological status. *Genome Med.* **7,** 49 (2015).
329 34. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome
330 assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
331 35. HannonLab. FASTX toolkit. *Cold Spring Harbor Laboratory, Cold Spring Harbor, NY* (2014).
332 36. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human
333 and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* **33,** D256-61 (2005).
334 37. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.*
335 **147,** 195–197 (1981).

336 38. Galson, J. D. *et al.* B-cell repertoire dynamics after sequential hepatitis B vaccination and
337 evidence for cross-reactive B-cell activation. *Genome Med.* **8,** 68 (2016).
338 39. Galson, J. *et al.* Analysis of B Cell Repertoire Dynamics Following Hepatitis B Vaccination in
339 Humans, and Enrichment of Vaccine-specific Antibody Sequences. *EBioMedicine* **2,** 2070–2079
340 (2015).
341 40. Greiff, V. *et al.* Quantitative assessment of the robustness of next-generation sequencing of
342 antibody variable gene repertoires from immunized mice. *BMC Immunol.* **15,** 40 (2014).
343 41. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain
344 sequence analysis tool. *Nucleic Acids Res.* **41,** (2013).
345 42. Dunbar, J. & Deane, C. M. ANARCI: Antigen receptor numbering and receptor classification.
346 *Bioinformatics* btv552 (2015). doi:10.1093/bioinformatics/btv552
347 43. Lefranc, M.-P. *et al.* IMGT unique numbering fro immunoglobulin and T cell receptor variable
348 domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* **27,** 55–77 (2003).
349 44. Eddy, S. R. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*
350 **3,** 114–120 (1995).
351 45. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11,** 653–5
352 (2014).
353 46. Banerjee, S. *et al.* Evaluation of a novel multi-immunogen vaccine strategy for targeting
354 4E10/10E8 neutralizing epitopes on HIV-1 gp41 membrane proximal external region. *Virology*
355 **505,** 113–126 (2017).
356 47. Bashford-Rogers, R. J. M. *et al.* Network properties derived from deep sequencing of human B-
357 cell receptor repertoires delineate B-cell populations. *Genome Res.* **23,** 1874–1884 (2013).
358 48. Bhiman, J. N. *et al.* Viral variants that initiate and drive maturation of V1V2-directed HIV-1
359 broadly neutralizing antibodies. *Nat. Med.* **21,** 1332–1336 (2015).
360 49. Bonsignori, M. *et al.* Maturation Pathway from Germline to Broad HIV-1 Neutralizer of a CD4-
361 Mimic Antibody. *Cell* **165,** 449–463 (2016).
362 50. Collins, A. M. *et al.* The mouse antibody heavy chain repertoire is germline-focused and highly
363 variable between inbred strains. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** S41–S52 (2015).
364 51. Corcoran, M. M. *et al.* Production of individualized v gene databases reveals high levels of
365 immunoglobulin genetic diversity. *Nat. Commun.* **7,** (2016).
366 52. Cui, A. *et al.* A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig
367 Sequencing Data. *J. Immunol.* **197,** 3566–3574 (2016).

368 53. Fisher, C. R. *et al.* T-dependent B cell responses to Plasmodium induce antibodies that form a
369 high-avidity multivalent complex with the circumsporozoite protein. *PLoS Pathog.* **13,** (2017).
370 54. Gupta, N. T. *et al.* Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig
371 Repertoire Sequencing Data. *J. Immunol.* **198,** 2489–2499 (2017).
372 55. Halliley, J. L. *et al.* Long-Lived Plasma Cells Are Contained within the CD19(-)CD38(hi)CD138(+)
373 Subset in Human Bone Marrow. *Immunity* **43,** 132–45 (2015).
374 56. Huang, J. *et al.* Identification of a CD4-Binding-Site Antibody to HIV that Evolved Near-Pan
375 Neutralization Breadth. *Immunity* **45,** 1108–1121 (2016).
376 57. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza
377 vaccination. *Sci. Transl. Med.* **5,** 171ra19 (2013).
378 58. Joyce, M. G. *et al.* Vaccine-Induced Antibodies that Neutralize Group 1 and Group 2 Influenza A
379 Viruses. *Cell* **166,** 609–623 (2016).
380 59. Khan, T. A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification
381 fingerprinting. *Sci. Adv.* **2,** e1501371–e1501371 (2016).
382 60. Levin, M. *et al.* Persistence and evolution of allergen-specific IgE repertoires during subcutaneous
383 specific immunotherapy. *J. Allergy Clin. Immunol.* **137,** 1535–1544 (2016).
384 61. Levin, M., Levander, F., Palmason, R., Greiff, L. & Ohlin, M. Antibody-encoding repertoires of
385 bone marrow and peripheral blood—a focus on IgE. *J. Allergy Clin. Immunol.* **139,** 1026–1030
386 (2017).
387 62. Li, X. *et al.* Comparative analysis of immune repertoires between bactrian Camel's conventional
388 and heavy-chain antibodies. *PLoS One* **11,** (2016).
389 63. Liao, H. X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*
390 **496,** 469–476 (2013).
391 64. Lindner, C. *et al.* Diversification of memory B cells drives the continuous adaptation of secretory
392 antibodies to gut microbiota. *Nat. Immunol.* **16,** 880–888 (2015).
393 65. Meng, W. *et al.* An atlas of B-cell clonal distribution in the human body. *Nat. Biotechnol.* **35,** 879–
394 886 (2017).
395 66. Menzel, U. *et al.* Comprehensive evaluation and optimization of amplicon library preparation
396 methods for high-throughput antibody sequencing. *PLoS One* **9,** (2014).
397 67. Mroczek, E. S. *et al.* Differences in the composition of the human antibody repertoire by B cell
398 subsets in the blood. *Front Immunol* **5,** 96 (2014).
399 68. Ota, M. *et al.* Regulation of the B cell receptor repertoire and self-reactivity by BAFF. *J. Immunol.*

400 **185,** 4128–36 (2010).
401 69. Palanichamy, A. *et al.* Immunoglobulin class-switched B cells form an active immune axis
402 between CNS and periphery in multiple sclerosis. *Sci. Transl. Med.* **6,** (2014).
403 70. Prohaska, T. A. *et al.* Massively Parallel Sequencing of Peritoneal and Splenic B Cell Repertoires
404 Highlights Unique Properties of B-1 Cell Antibodies. *J. Immunol.* **200,** 1702–1717 (2018).
405 71. Rubelt, F. *et al.* Individual heritable differences result in unique cell lymphocyte receptor
406 repertoires of naïve and antigen-experienced cells. *Nat. Commun.* **7,** (2016).
407 72. Zhu, J. *et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation
408 sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. U. S. A.* **110,**
409 6470–5 (2013).
410 73. Stern, J. N. H. *et al.* B cells populating the multiple sclerosis brain mature in the draining cervical
411 lymph nodes. *Sci. Transl. Med.* **6,** (2014).
412 74. Sundling, C. *et al.* Single-Cell and Deep Sequencing of IgG-Switched Macaque B Cells Reveal a
413 Diverse Ig Repertoire following Immunization. *J. Immunol.* **192,** 3637–3644 (2014).
414 75. Tipton, C. M. *et al.* Diversity, cellular origin and autoreactivity of antibody-secreting cell
415 population expansions in acute systemic lupus erythematosus. *Nat. Immunol.* **16,** 755–765
416 (2015).
417 76. Turchaninova, M. A. *et al.* High-quality full-length immunoglobulin profiling with unique
418 molecular barcoding. *Nat. Protoc.* **11,** 1599–1616 (2016).
419 77. Vander Heiden, J. A. *et al.* Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis
420 Patients Revealed through Deep Sequencing. *J. Immunol.* **198,** 1460–1473 (2017).
421 78. VanDuijn, M. M., Dekker, L. J., van IJcken, W. F. J., Sillevis Smitt, P. A. E. & Luider, T. M. Immune
422 repertoire after immunization as seen by next-generation sequencing and proteomics. *Front.*
423 *Immunol.* **8,** (2017).
424 79. Vergani, S. *et al.* Novel method for high-throughput full-length IGHV-D-J sequencing of the
425 immune repertoire from bulk B-cells with single-cell resolution. *Front. Immunol.* **8,** (2017).
426 80. Wesemann, D. R. *et al.* Microbial colonization influences early B-lineage development in the gut
427 lamina propria. *Nature* **501,** 112–115 (2013).
428 81. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep
429 sequencing. *Science* **333,** 1593–602 (2011).
430 82. Wu, X. *et al.* Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic
431 HIV-1 infection. *Cell* **161,** 480–485 (2015).

432 83. Wu, Y. C. B. *et al.* Influence of seasonal exposure to grass pollen on local and peripheral blood IgE
433 repertoires in patients with allergic rhinitis. *J. Allergy Clin. Immunol.* **134,** 604–612 (2014).
434 84. Zhou, T. *et al.* Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in
435 14 donors. *Cell* **161,** 1280–1292 (2015).
436 85. Zhu, J. *et al.* Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454
437 pyrosequencing and bioinformatics. *Front. Microbiol.* **3,** (2012).
438 86. Zhu, J. *et al.* De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-
439 generation sequencing of B-cell transcripts. *Proc. Natl. Acad. Sci.* **110,** E4088–E4097 (2013).
440 87. Schroeder Jr, H. W. Similarity and divergence in the development and expression of the mouse
441 and human antibody repertoires. *Dev. Comp. Immunol.* **30,** 119–135 (2006).
442 88. Zemlin, M. *et al.* Expressed murine and human CDR-H3 intervals of equal length exhibit distinct
443 repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol.*
444 *Biol.* **334,** 733–749 (2003).
445 89. Olimpieri, P. P., Marcatili, P. & Tramontano, A. Tabhu: Tools for antibody humanization.
446 *Bioinformatics* **31,** 434–435 (2014).
447 90. Lowe, D. *et al.* Aggregation, stability, and formulation of human antibody therapeutics. in 41–61
448 (2011). doi:10.1016/B978-0-12-386483-3.00004-5
449 91. Lauer, T. M. *et al.* Developability index: A rapid in silico tool for the screening of antibody
450 aggregation propensity. *J. Pharm. Sci.* **101,** 102–115 (2012).
451