Synthetic standards combined with error and bias correction improves the accuracy and quantitative resolution of antibody repertoire sequencing in human naïve and memory B cells

Simon Friedensohn*[1], John M. Lindner*[2], Vanessa Cornacchione[2], Mariavittoria Iazeolla[2], Enkelejda Miho[1,3], Andreas Zingg[1], Simon Meng[1], Elisabetta Traggiai[2], and Sai T. Reddy[1]

[1]*Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
[2]*Novartis Institutes for BioMedical Research, Basel, Switzerland*
[3]*aiNET GmbH, c/o ETH Zürich, Basel, Switzerland*

*equal contribution

Correspondence: elisabetta.traggiai@novartis.com and sai.reddy@ethz.ch

ABSTRACT

High-throughput sequencing of immunoglobulin repertoires (Ig-seq) is a powerful method for quantitatively interrogating B cell receptor sequence diversity. When applied to human repertoires, Ig-seq provides insight into fundamental immunological questions, and can be implemented in diagnostic and drug discovery projects. However, a major challenge in Ig-seq is ensuring accuracy, as library preparation protocols and sequencing platforms can introduce substantial errors and bias that compromise immunological interpretation. Here, we have established an approach for performing highly accurate human Ig-seq by combining synthetic standards with a comprehensive error and bias correction pipeline. First, we designed a set of 85 synthetic antibody heavy chain standards (*in vitro* transcribed RNA) to assess correction workflow fidelity. Next, we adapted a library preparation protocol that incorporates unique molecular identifiers (UIDs) for error and bias correction which, when applied to the synthetic standards, resulted in highly accurate data. Finally, we performed Ig-seq on purified human circulating B cell subsets (naïve and memory), combined with a cellular replicate sampling strategy. This strategy enabled robust and reliable estimation of key repertoire features such as clonotype diversity, germline segment and isotype subclass usage, and somatic hypermutation (SHM). We anticipate that our standards and error and bias correction pipeline will become a valuable tool for researchers to validate and improve accuracy in human Ig-seq studies, thus leading to potentially new insights and applications in human antibody repertoire profiling.

INTRODUCTION

Adaptive immune responses are governed by cooperative interactions between B and T lymphocytes upon antigen recognition. A hallmark of these cells is the somatic generation of clonally unique antigen receptors during primary lymphocyte differentiation. In particular, B cell antigen receptors (BCRs, and their analogous secreted form, antibodies) result from rearrangement of the germline-encoded variable (V), diversity (D, heavy chain only), and joining (J) gene segments. V(D)J recombination in B cells creates a highly complex receptor population (generally interchangeably referred to as BCR, antibody, or immunoglobulin (Ig) repertoires), which matures upon antigen experience to produce the more targeted, high-affinity memory BCR network. In-

depth and accurate characterization of these repertoires provides valuable insight into the generation and maintenance of immunocompetency, which can be used to monitor changes in immune status, and to identify potentially reactive clones for therapeutic or other uses. Due to rapid technological advances, high-throughput sequencing of Ig genes (Ig-seq) has become a major approach to catalog the diversity of antibody repertoires (1-3). Ig-seq applied to human B cells has potential in a variety of applications (4), particularly in antibody drug discovery (5-7), diagnostic profiling for vaccines (8, 9), and biomarker-based disease detection (10, 11). Additionally, Ig-seq is enabling a more comprehensive understanding of basic human immunobiology, such as B cell clonal distribution across physiological compartments in health and disease (12, 13).

A major challenge in Ig-seq is the requirement of accurate and high-quality datasets. Several current library preparation protocols are based on target enrichment from genomic DNA or mRNA (14). For example, the conversion of mRNA (more commonly used due to transcript abundance and isotype splicing) into antibody sequencing libraries relies on a number of molecular reagents and amplification steps (e.g. reverse transcriptase, multiplex primer sets, PCR), which potentially introduce errors and bias. Due to the highly polymorphic nature of repertoires especially from affinity-matured memory B cells and plasma cells, it becomes essential to determine if such technical noise occurs at non-negligible rates, as this could alter quantitation of critical repertoire features such as clonal frequencies, germline gene usage, and somatic hypermutation (SHM) (14, 15). One way to address this is through the use of synthetic control standards, for which the sequence and abundance is known prior to sequencing, thus providing a means to assess quality and accuracy (16). Several examples of standards have been presented for Ig-seq; Shugay et al. sequenced libraries prepared from a small polyclonal pool of B and T lymphocyte cell lines, and observed nearly 5% erroneous reads, resulting in approximately 100 false-positive variants per clone (17). Recently, Khan et al. developed a set of synthetic RNA (*in vitro* transcribed) spike-in standards based on mouse antibody sequences, which were used to show that a substantial amount of errors and bias are introduced during multiplex-PCR library preparation and sequencing (18).

Various experimental and computational workflows exist to mitigate the effects of errors and bias in Ig-seq. One of the most advanced and powerful strategies is to prepare libraries with the incorporation of random and unique molecular identifiers (UIDs, also commonly referred to as UMIs or molecular barcodes). Following sequencing, error correction can be performed by clustering and consensus building of reads that share the same UID; reads sharing the same UID are assumed to be derived from the same original mRNA/cDNA molecule (19). Furthermore, bias correction for cDNA abundance can be performed by counting the number of UIDs (instead of total reads) (20, 21). Several iterations of UID-tagging have been developed for Ig-seq, such as UID labeling during first- and second-strand cDNA synthesis (22), UID addition during RT template switching (23), and so-called "tagmentation" of UID-labelled amplicons (24). Recently, we developed an innovative strategy to add UIDs both during first-strand cDNA synthesis as well as multiplex-PCR amplification; this protocol, known as molecular amplification fingerprinting (MAF), results in comprehensive error and bias correction of mouse antibody repertoires (18).

Here, we describe an experimental-computational approach to generate highly accurate human Ig-seq data. We first designed a comprehensive set of synthetic standards based on human antibody heavy chain variable (IGHV) sequences: a total of 85 *in vitro* transcribed RNA standards, each with a unique complementarity determining region 3 (CDR3) sequence and covering nearly the entire set

of productive human Ig heavy chain (IgH) germline (IGHV) gene segments. We used these synthetic standards to quantify the impact of errors and bias introduced during multiplex-PCR library preparation, and the robustness with which our previously developed method for UID addition by MAF could correct these artifact sequences. Finally, we implemented MAF-based error and bias correction on human B cell subsets (naïve and memory), which enabled us to make accurate clonal diversity estimates and quantify divergent repertoire features across B cell compartments.

RESULTS

Design of a comprehensive set of human synthetic standards

Our previously established set of murine synthetic antibody standards contained 16 unique clones (CDR3s) covering 7 IGHV gene segments (out of more than 140 annotated murine IGHV gene segments) (18). For our human standards, we developed a more comprehensive set consisting of 85 clones encompassing nearly the entire germline IGHV repertoire. The most commonly used repository for human germline segments is the International ImMunoGenetTics Database (IMGT), which has annotated 61 IGHV alleles as functional or having an open reading frame (25). After filtering out paralogs and selecting only gene segments that have been found in productive rearrangements (2), we chose 48 IGHV gene segments as the basis for our standards (**Table S1**). Each standard contained the following elements (5' to 3'): (i) a conserved non-coding region, (ii) ATG start codon and a leader peptide sequence spliced to its respective IGHV gene segment, (iii) a synthetic CDR3 sequence, (iv) a germline IgH J (IGHJ) gene segment, (v) a non-coding synthetic sequence identifier (for the separation of standards from biological sequences), (vi) a partial segment of the constant region from isotypes IgM, IgG, and IgA (**Figure 1A**). This design allows amplification of our synthetic controls with a variety of PCR primer sets. Notably, for control singleplex-PCR experiments, all standards can be amplified by a single forward primer (targeting the conserved 5' non-coding region) and a single reverse primer (targeting one of the isotype constant regions). Since IGHV gene segment usage has been reported to be non-uniform (2, 26), we selected the most abundant segments for use in multiple standards (**Figure 1B, Table S1**).

All standards carry a unique CDR3 sequence, which visually aids the analysis of sequencing results. Furthermore, all clones were designed to be resilient against sequencing and PCR errors: at least 9 specific nucleotide (nt) deletions, insertions, and/or mutations are needed in order to turn one CDR3 nt sequence into another (**Figure 1B**). For our experiments, synthetic standard genes were *in vitro* transcribed to RNA and subsequently reverse transcribed to cDNA. We measured individual cDNA molecules by digital droplet PCR (ddPCR) and capillary electrophoresis. Standards were then pooled in a non-uniform concentration distribution and maintained as a master stock (**Table S1**).

Human Ig-seq library preparation using the MAF protocol

We adapted our previously described library preparation protocol for murine antibody repertoires to be compatible for human Ig-seq (18). This protocol is based on targeted amplification via RT of RNA to first-strand cDNA, followed by two PCR amplification steps (27, 28), the first of which uses a forward multiplex primer set targeting the IGHV framework region 1 (FR1). Each step also

133 incorporates fragments of Illumina sequencing adapters (IA), such that the final product of the
134 workflow is already compatible with the Illumina sequencing platform (**Figure 2A**).

135 Importantly, our library preparation protocol used primers incorporating random-nucleotide UIDs,
136 thus enabling MAF-based error and bias correction. A reverse-UID (RID) with theoretical diversity
137 up to $2x10^7$ unique sequences is present in the RT primer (between the Ig constant region-specific
138 and partial IA regions), and a forward-UID (FID with additional diversity of approximately $7x10^5$
139 unique sequences) is present on the forward multiplex primer set (between the FR1-specific and
140 partial IA regions) used in the first PCR reaction. Such high diversity among RIDs is necessary in
141 order to prevent tagging of different cDNA molecules with the same barcode. Our multiplex
142 forward primer set was designed to target all IMGT-annotated IGHV gene segments (**Figure 2B**).
143 To compromise between maintaining similar amplicon length across gene segment families and
144 creating thermodynamically equivalent oligonucleotides, we placed the primers at or near the
145 beginning of each FR1, resulting in a melting temperature range of 57°C to 63°C (**Figure 2C**). This
146 range and the accompanying (unavoidable) variability in GC content have been shown to
147 potentially cause differences in amplification efficiencies, which in turn leads to a biased
148 representation of segment usage frequencies in Ig-seq data. Our workflow aimed to solve this
149 problem in two ways: first, since the RID labels cDNA at the single molecule level, we are able to
150 resolve the number of molecules by counting the number of RIDs instead of raw reads. Second, by
151 using the FIDs on our forward primers, we are able to further normalize our molecular count, since
152 Ig genes preferentially amplified by our primer set should show a higher ratio of FIDs to RIDs.
153 Additionally, the RIDs can be used to correct for errors introduced during PCR and the sequencing
154 process itself by grouping sequencing reads based on their RID, then correcting diverging nt
155 positions by generating a consensus sequence (majority voting scheme). This is especially useful in
156 Ig-seq when attempting to distinguish true SHM variants from erroneous sequences.

157 **Combining standards with MAF to correct errors and bias in Ig-seq**

158 To evaluate the extent of errors and bias present in human Ig-seq data, standards were mixed
159 (spiked-in) with cDNA prepared from circulating purified human B cells. In total, we sequenced 28
160 independently prepared libraries and annotated them with a custom aligner (18, 29). Prior to
161 alignment, reads were either kept as uncorrected (raw) reads or were corrected using our MAF
162 pipeline that takes into account RID and FID information. In this way, we could directly compare
163 the number of erroneous sequences produced in uncorrected vs. MAF-corrected datasets. Clonal
164 assignment of uncorrected reads produced many erroneous CDR3 amino acid (a.a.) variants
165 (sequences with at least 1 a.a. difference from the nearest standard control sequence); for example,
166 in a dataset with 100,000 aligned reads, there was a median value of 23 errors per clone (**Figure**
167 **3A**). The number of erroneous variants produced showed a clear correlation with the individual
168 abundance of each clone within the master stock (r = 0.89). When taking the entire VDJ nt
169 sequence into account, an even greater number of erroneous variants were observed (≥1 nt
170 difference from the standard sequence) (**Figure 3B**). We observed a median value of 118 erroneous
171 nt variants per standard (per 100,000 aligned sequences). Again, the number of erroneous variants
172 exhibited a clear correlation with clone abundance (r = 0.90). However, we did not observe any
173 significant trend linking IGHV family to the error rate (F-Test on full and reduced linear model, p =
174 0.083).

175 After performing error correction with our MAF pipeline, there was a dramatic reduction of CDR3
176 and VDJ errors: we observed a median value of 0 and 1 error per clone, respectively. Across all
177 datasets, we remove an average of 94.2% CDR3 a.a. and 97.4% VDJ nt erroneous variants (**Figure**
178 **3A-B**). For example, prior to error correction the standard "CARGINGERALEW" (from dataset
179 Donor 1, IgG aliquot 1, see **Table S3**) displayed 39 additional CDR3 a.a. variants and 217
180 additional VDJ nt variants (**Figure 3D**). After error correction, we retain only the correct CDR3 a.a.
181 sequence and only one additional (erroneous) nt variant. With our current filtering criteria (see
182 Methods), we observed 16 instances in which we removed a standard control sequence that was
183 present in the raw data. However, in the vast majority of cases, we kept the standard when it was
184 observed in the raw data (2,321 instances). In only 43 instances, a standard sequence was either too
185 low in abundance or too frequently mutated to be annotated in either the raw or error-corrected
186 datasets.

187 In order to assess potential biases introduced by library preparation we prepared control libraries
188 containing only the pool of synthetic standards (from the master stock). The libraries were
189 generated in the same manner as the described MAF protocol, with the exception that in the first
190 PCR step, instead of using a multiplex forward primer set, a single forward primer targeting the
191 conserved 5' non-coding region (singleplex-PCR) was used. Ig-seq on these samples allowed us to
192 establish a baseline for pipetting accuracy by comparing the obtained standard frequencies from the
193 singleplex-PCR against the expected frequencies based on our pooling scheme: this yielded an R^2
194 of 0.88 and average mean squared error (MSE) of 0.29 ± 0.02% (**Figure S1A**). These values
195 indicate that only small systematic deviations occurred, most likely due to minor pipetting error.
196 Next, we compared standard frequencies (expected relative concentration) with frequencies
197 generated in our previous multiplex-PCR libraries, both with and without MAF correction (**Figure**
198 **3C**). On uncorrected data, the multiplex-PCR libraries achieved an R^2 of 0.84 with an average MSE
199 of 0.34 ± 0.06%, which is significantly worse than the value obtained by singleplex-PCR (Student's
200 t-test $p = 0.008$). After error and bias correction on these same datasets, the correlation improved to
201 an R^2 of 0.89 and an average MSE of 0.28 ± 0.08%. While MAF-corrected MSE values show no
202 significant difference to the singleplex-PCR libraries (Student's t-test $p = 0.49$), they do highlight a
203 significant improvement over the uncorrected data (paired Student's t-test, $p = 0.0007$).

204 **Impact of MAF error correction on human B cell repertoires**

205 Next, we analyzed the impact of MAF on the BCR repertoires of B cells isolated from the
206 peripheral blood of three healthy donors. We used flow cytometry sorting and a gating strategy to
207 select for $CD27^-$ IgM^+ (naïve) and $CD27^+$ IgG^+ (memory) B cell populations (**Figure 4A**). Across
208 all donors, the fraction of $CD19^+$ peripheral blood B cells was 16-29% for $CD27^-$ IgM^+ and 5-9%
209 for $CD27^+$ IgG^+. Importantly, each donor population was split into 4-5 separate aliquots containing
210 200,000 cells each (cellular replicates) prior to cell lysis. Total RNA was extracted and RT for
211 cDNA synthesis was performed independently in order to prevent the mixing of transcripts across
212 cellular replicates. The cDNA of synthetic standards (from the master stock) was then mixed with
213 the B cell cDNA, and corresponding molecular quantities were measured by ddPCR (**Figure 4B,**
214 **Table S3**). All cDNA libraries were then processed into libraries using the MAF protocol (**Figure**
215 **2A**) and subjected to Ig-seq.

216 A simple global analysis of Ig-seq data revealed that diversity measurements were dramatically
217 exaggerated, as the number of unique antibody sequence variants obtained from the raw,

218 uncorrected data often exceeded both the number of cells and the number of total cDNA transcripts
219 in a given aliquot. Following error correction by MAF, the variant count returned to ranges that are
220 physically possible, thereby highlighting the importance of proper error correction and quality
221 control when globally determining repertoire diversity (**Figure 4B**). We further examined the
222 influence of erroneous variants on CDR3 clonotype analysis. In order to identify clonotypes, we
223 used hierarchical clustering (30) based on sequences sharing the following features: identical IGHV
224 and IGHJ gene segment usage, identical CDR3 length, and a CDR3 a.a. similarity of at least 80% to
225 one other sequence in the given clonotype. When performing such an analysis on uncorrected data,
226 clonotypes contained an artificially high number of distinct clones (**Figure 4C**, left tree). Here, the
227 IgG-derived clonotype with the consensus CDR3 of 'CARAAGSQYYMDVW' (from the same
228 sample and IGHV gene segment used by the standard shown in **Figure 3D**) contains 249 unique nt
229 variants and 70 unique a.a. sequences. After MAF-based error correction, only 15 nt variants and 6
230 distinct CDR3 a.a. sequences remained. It is worthy to note that although both the standard
231 sequence (**Figure 3D**) and the biological clonotype (**Figure 4C**) had a large number of CDR3 a.a.
232 variants in uncorrected data, after error correction only the biological clonotype retained multiple
233 a.a. variants, suggesting these may be true variants generated *in vivo* by SHM. This general trend of
234 each IgG memory B cell-derived clonotype to contain more variants relative to antigen
235 inexperienced IgM-expressing B cells was clear across our biological data sets (**Figure 5A**).

236 **Clonal diversity measurements of human B cell repertoires after error and bias correction**

237 After establishing the value of performing MAF error correction on biological repertoires, we next
238 focused on determining the clonotype diversity present in each B cell sample. First, we determined
239 the overlap of clonotypes present in each cellular replicate (**Figure 6A**). Notably, we observed an
240 overlap of several clonotypes in the $CD27^{-}IgM^{+}$ subset; this overlap was unexpected given that this
241 subset should be highly enriched for naïve B cells, which by definition are not antigen experienced
242 or clonally expanded, and should therefore be mostly unique (not present in multiple replicates).
243 For each donor in the $CD27^{-}IgM^{+}$ subset, 1- 2% of all clonotypes were present in at least one other
244 cellular replicate. In the $CD27^{+}IgG^{+}$ subset, clonotypes shared between at least two cellular
245 replicates were nearly tenfold more frequent (12-15%), which was expected given that this
246 population is comprised of antigen experienced, clonally expanded memory B cells. Another
247 observation discordant with the expected naïve B cell properties of the $CD27^{-}IgM^{+}$ subset was that
248 overlapping clonotypes (in donors 1 and 3) were significantly more likely to have acquired
249 mutations (**Figure 6B**), which are not a typical feature of naïve B cells. In comparison, over 90% of
250 all $CD27^{+}IgG^{+}$ (overlapping and non-overlapping) clonotypes possessed at least one SHM, an
251 expected observation in a memory subset.

252 The high amount of overlap within the $CD27^{+}IgG^{+}$ B cell replicates of each donor allowed us to use
253 established population diversity estimation techniques to calculate clonal diversity (31). Rarefaction
254 curves were generated and estimates were extrapolated as a function of real and predicted cellular
255 replicates (**Figure 6C**). The asymptote was determined by the standard form of the Chao2 estimator
256 and yielded the following values for clonotype numbers: donor 1 = 164,268 ± 2,365, donor 2 =
257 38,034 ± 1,302, and donor 3 = 76,904 ± 1,409. Since the 95% confidence intervals for the three
258 donors did not overlap, we could also infer that the size of each donor's repertoire at the collection
259 time point was significantly different. This analysis indicates that we would need to sample at least
260 tenfold more cellular replicates in order to observe > 90% of all clonotypes; however the first five

261 samples analyzed here were sufficient to observe > 25% of the clonotypic memory repertoire. We
262 also generated rarefaction and extrapolation curves rescaled to the RID count (**Figures S2A and**
263 **S2B**). In the case of the $CD27^-IgM^+$ repertoire data, while asymptotic curves could be generated, a
264 diversity estimation is impractical. This is because plotting the observed numbers of newly
265 discovered clonotypes for each additional RID and donor shows that the number of newly
266 discovered clonotypes in the $CD27^-IgM^+$ dataset continues to grow over the observed range,
267 whereas the number of new clonotypes starts to converge at approximately 20,000 RIDs for
268 $CD27^+IgG^+$ repertoires (**Figure S2C**).

269 **Divergent features of $CD27^-IgM^+$ and $CD27^+IgG^+$ repertoires**

270 After pooling all clonotypes (expanded and unique to a single cellular replicate) for each donor, we
271 globally characterized sequences of the naïve $CD27^-IgM^+$ and memory $CD27^+IgG^+$ subsets. First,
272 we determined the SHM count (nt) of each clone with respect to its nearest germline IGHV and
273 IGHJ gene segment sequence. The median values of SHM for the $CD27^-IgM^+$ repertoires were
274 zero, which was to be expected for a naïve B cell subset. In contrast, the median values for
275 $CD27^+IgG^+$ repertoires were 20-24 mutations per clone (**Figure 5B**). It is widely appreciated that
276 human heavy chain CDR3 sequences are much longer than their murine counterparts, which we
277 also observed here, with a slight (but consistent across donors) variation between naïve and memory
278 B cells (**Figure 5C**). Interestingly, the IGHV gene segment family usage correlated with B cell
279 subset. The $CD27^+IgG^+$ repertoires across all donors were relatively enriched for IGHV1 and
280 IGHV3 gene segment family members, whereas the relative share of the IGHV4 gene segment
281 family was larger in the $CD27^-IgM^+$ repertoires (**Figure 7A**). We validated these observations
282 quantitatively using linear discriminant analysis (LDA) fitted on the centered log ratio (CLR)-
283 transformed frequencies of each cellular replicate (**Figure S3**). The LDA classifier was fit on
284 different splits of the data (based on two of the donors) and used to predict a holdout set (based on
285 the remaining donor). This showed that the fitted classifier in each instance is highly predictive of
286 the remaining aliquots and that prediction is robustly driven by the relative abundance of IGHV4
287 segment family usage in the $CD27^-IgM^+$ repertoires and the IGHV1, 2, and 3 families in the
288 $CD27^+IgG^+$ repertoires (**Figure S3A-C**). Next, we utilized LDA to perform dimensionality
289 reduction of all data points to into a single one-dimensional axis; again, the most important
290 components were the relative abundance of IGHV3 and IGHV4 gene segment families (**Figure**
291 **S3D**).

292 Next, we leveraged the ability of our reverse primer to distinguish among IgG subclasses (**Figure**
293 **7B**). The majority of sequences mapped either to IgG1 (40-66%), IgG2 (23%-36%), or IgG3 (23%-
294 36%), whereas IgG4 sequences were extremely rare, observed solely in donor 3 (0.3%). Finally, we
295 compared the $CD27^-IgM^+$ and $CD27^+IgG^+$ repertoires of each donor to determine the clonotype
296 overlap of each B cell subset and isotype. Strikingly, the observed overlap was very small, with
297 only 269 shared clonotypes for donor 1, 30 shared clonotypes for donor 2, and 215 clonotypes for
298 donor 3 (**Figure 7C**). In each donor, these represented less than 0.5% of identified clonotypes.
299 Closer examination revealed that clonotypes shared between the $CD27^-IgM^+$ and $CD27^+IgG^+$
300 subsets were also significantly enriched for intraclonal variants (SHM in CDR3) in one of the two
301 populations (**Figure 7D and Figure S4**). Furthermore, we could see that clonotypes with multiple
302 IgM variants were also shared specifically among IgM cellular replicates (**Figure 7D**, contingency
303 tables). This intraclonal variant bias was not limited to heavy chain isotype: IgG clonotypes with ≥

304 5 intraclonal variants also exhibited subclass composition skewing toward the IgG1/2 or IgG1/3
305 axis, but rarely at proportions similar to the overall IgG subclass distribution (**Figure 7E**).

306

307 **DISCUSSION**

308 Ig-seq is becoming an essential tool for the quantitative analysis of antibody repertoire diversity and
309 distribution. However, similar to other areas of high-throughput sequencing, Ig-seq also suffers
310 from technical errors and bias; thus, standardized experimental and analytical methods that increase
311 the validity of immunological interpretations must be developed. Here, we establish a
312 comprehensive set of synthetic standards which, when combined with UID-labeling and MAF-
313 based error and bias correction, results in highly accurate antibody repertoire data. By applying this
314 approach to human B cell subsets, we gain unique insights into repertoire features such as clonal
315 diversity, germline gene usage, SHM, and clonal history.

316 The synthetic standards developed here allowed quantitative interrogation of several accuracy-
317 related features in Ig-seq. One major observation was that raw uncorrected data has a high number
318 of erroneous variants, found both within the clonotype-defining CDR3 and across the entire VDJ
319 region (**Figure 3A, B**). The number of false-positive variants correlated with the abundance of each
320 standard; this is of particular concern because high frequency, clonally expanded B cells are often
321 correlated with antigen specificity and thus important for biological interpretations (32). However,
322 when applying our MAF-error correction protocol, we were able to remove nearly all erroneous
323 CDR3 and VDJ variants (94-97%); this correction was robust even for high-frequency standards
324 where the number of erroneous variants was especially high. Errors not removed by MAF could
325 potentially be addressed with more stringent filtering criteria (e.g. read number cutoffs); however,
326 this may come at the cost of reducing overall dataset size and removing legitimate intraclonal
327 variants in biological samples.

328 Another aspect we quantified with our standard pool was the impact of multiplex primer sets, which
329 have been shown to introduce substantial bias during library preparation (18, 33). By designing our
330 standards with a 5' conserved singleplex region (**Figure 1A**), we were able to directly compare Ig-
331 seq data from libraries (on the same master stock) prepared by singleplex-PCR vs. multiplex-PCR.
332 Our newly designed FR1-targeting multiplex primer set (**Figure 2B, C**) demonstrated a relatively
333 strong correlation with singleplex-PCR data (R^2 = 0.84) (**Figure 3C**). However, by performing
334 multiple MAF error and bias correction steps, the correlation was improved by an additional 7%
335 (**Figure 3D)**. The remaining variability does not appear to be restricted to a particular IGHV-gene
336 family, indicating there is little systematic bias with respect to homologous sequences within the
337 standard pool. MAF therefore represents an essential step in eliminating technical artifacts from
338 human Ig-seq workflows, as it is able to generate data that closely mirrors that of the original
339 sample. In future applications, these synthetic standards could be a critical asset for evaluating
340 newly designed primer sets, library preparation protocols, or implementing new error and bias
341 correction pipelines.

342 Having established a comprehensive set of synthetic standards and a validated error and bias
343 correction pipeline, we were able to perform several analyses on human B cell repertoires with
344 greater confidence in the accuracy and quantitative resolution of the Ig-seq data. A simple approach

345 to estimating repertoire diversity is to calculate the number of unique antibody sequences as a
346 fraction of total transcript (cDNA) input. However, performing this analysis on our samples
347 suggests that the $CD27^{+}IgG^{+}$ memory B cell compartment is significantly more diverse than the
348 naïve $CD27^{-}IgM^{+}$ B cells (82% vs. 35% unique nt variants, respectively, averaged across donors
349 and cellular aliquots, Student's t-test $p < 10^{-4}$). This is potentially due to sample size variability with
350 respect to the number of transcripts and our ability to oversample smaller libraries. Critically, when
351 using bulk-sorted cells with UID labeling, it is not possible to discriminate between transcript
352 copies that are identical because they came from the same lysed cell, and those which are identical
353 because they represent two distinct, clonally related B cells. Thus, by biological subsampling
354 through cellular replicates, we ensured that clonotypes observed in multiple samples must come
355 from distinct, clonally related B cells, thereby providing an effective solution for estimating clonal
356 diversity.

357 Applying computational approaches from ecology (31) to our biological subsampling strategy, we
358 attempted to estimate the number of unique clonotypes in a given antibody repertoire. The $CD27^{-}$
359 IgM^{+} B-cell subset did not show substantial clonotype overlap among cellular aliquots (**Figure 6A**).
360 As it is commonly assumed (and typically the case as shown in **Figure 5A**) that each newly
361 generated B cell is a unique clone, the size of the naïve repertoire in the human peripheral blood
362 would be equal to the total number of naïve B cells, in the range of 10 to 30 million. While it is
363 improbable to sample this subset in its entirety, and its diversity is also too high to estimate based
364 on the cell numbers obtained here, our observations are consistent with this model, since each
365 additional cellular replicate produced overwhelmingly unique sequences. One donor did show an
366 unexpected presence of overlapping sequences (shared clonotypes) across IgM cellular replicates
367 (**Figure 6A**, donor 1); these clonotype sequences were significantly enriched for SHM (**Figure 6B**),
368 suggesting the possible presence of an antigen-experienced B cell subset within $CD27^{-}IgM^{+}$
369 population, and highlighting the need for improved characterization of the heterogeneity within
370 circulating human B cell subsets. In the $CD27^{+}IgG^{+}$ B cell subset, we observed substantially more
371 overlap across cellular replicates, which was expected given that memory-enriched B cells would
372 have experienced antigen and undergone clonal expansion. By extrapolating the numbers of
373 additional uniquely observed clonotypes with each subsequent cellular aliquot, we were able to
374 predict the clonotype size of the peripheral $CD27^{+}IgG^{+}$ B-cell repertoire to be on the order of 10^{5}
375 (**Figure 6C**). Indeed, rough estimates of the number of antigen-specific clonotypes generated by a
376 single immune response (≈100, a number in line with what has been described regarding serum
377 antibody clonotypes (34)) and the number of structurally distinct pathogens against which an
378 individual has mounted a response (≈1000, a generous estimate given work showing that
379 worldwide, individuals have on average a serological history against less than 100 viral species
380 (35)) suggest that a memory repertoire of this size could reasonably protect against latent infection
381 and/or subsequent antigen encounter.

382 We observed a clear shift in IGHV segment family usage from the naïve to the memory BCR
383 repertoire (**Figure 7A**). Consistently observing this reshaping in three independent healthy donors,
384 and comparing to our standard controls to exclude the possibility of biased amplification, we can
385 conclude that it is a genuine phenomenon. Relatively more abundant IGHV1 and less abundant
386 IGHV4 segment usage in IgG memory B cells has been previously observed in one three-donor
387 cohort (1) but not in another which pooled sequences from both class-switched and IgM-memory

388 cells (36), underscoring the importance of experimental design and accurate bias correction in
389 antibody repertoire analysis.

390 Our Ig-seq workflow also allowed us to unambiguously assign IgG antibody sequences to their
391 appropriate subclass, offering further insight into patterns of class-switch recombination present in
392 memory-enriched B cells. While plasma cell-secreted IgG proteins in human serum are present at
393 ratios of approximately 14:8:1:1 (for IgG1:2:3:4, respectively (37)), $CD27^{+}IgG^{+}$ B cells showed a
394 distribution of approximately 5:3:1 for IgG1:2:3, with a nearly complete absence of IgG4 (**Figure**
395 **7B**). Cole et al. similarly observed a lack of IgG4 heavy chains in a single donor but described an
396 enrichment of IgG2 relative to IgG1 and IgG3 (24). The abundance of $IgG3^{+}$ B cells relative to its
397 presence in the serum seen here indicates IgG3 may play a more important role in maintaining the
398 reactive memory response compared to the protective memory response provided by serum IgG.
399 Notably, the IgG3 locus is the most proximal, and thereby the most plastic of the human IgG
400 subclasses; that is, an $IgG3^{+}$ B cell still retains the capacity to class-switch to any of the remaining
401 three IgG subclasses, whereas IgG1, IgG2, and IgG4 cannot return to any of the previous states.
402 Similar to these findings, a flow cytometry-based investigation has also found healthy human
403 donors to have low frequencies of IgG4-expressing circulating memory B cells (38).

404 With new daily production and relatively rapid turnover of naïve B cells, it was not unexpected to
405 see little overlap of clonotypes between the intradonor $CD27^{-}IgM^{+}$ and $CD27^{+}IgG^{+}$ populations
406 (**Figure 7C**). An interesting finding was that for clonotypes present in both B cell subsets,
407 intraclonal variation was largely restricted to one of the two isotypes. Assuming that clonotype
408 overlap among $CD27^{-}IgM^{+}$ cellular replicates represents the presence of antigen-experienced,
409 clonally expanded B cells, this suggests that the antigen specificity of antibody variable domains
410 may to some extent be influenced by the downstream constant regions, which has been observed
411 functionally for small cohorts of human and murine IgG and IgA antibodies (39). Notably, we also
412 observe similar clonal restriction within IgG clonotypes with respect to heavy chain subclass
413 (**Figure 7E**). This may be driven by the type of antigen and the nature of the elicited immune
414 response, or governed by physical constraints as we suggest for the differences between the IgM
415 and IgG repertoires. A larger scale functional study, including IgM sequences, could provide crucial
416 support for this model, which would shed new light on the role of Ig isotypes and subclasses on B
417 cells in the post-antigen encounter setting.

418 **FIGURE LEGENDS**

419

420 **Figure 1.** A comprehensive set of human synthetic spike-in standards for Ig-seq. **(A)** Schematic
421 showing the prototypical spike-in with the following regions (5' to 3'): a conserved non-coding
422 region, ATG start codon, IGHV region with FR1 specific for multiplex-PCR, IGHJ regions, non-
423 coding synthetic sequence identifier (specific for ddPCR probes), and downstream heavy chain
424 constant domain sequences (IGHG, IGHM, IGHA) containing primer binding sites used for cDNA
425 synthesis. Each spike-in contains a complete VDJ open reading frame, including nucleotides
426 upstream of the ATG start codon, and downstream constant domain sequences containing primer
427 binding sites used for sample cDNA synthesis. **(B)** Pairwise comparisons based on a.a. Levenshtein
428 edit distance of all 85 standard CDR3 sequences. The germline IGHV and IGHJ segment family
429 usage and IgG subclasses are denoted. 39 spike-ins contain rationally designed nt SHM (black
430 circles) across the IGHV regions.

Figure 2. Library preparation of immunoglobulin (Ig) heavy chain genes for high-throughput sequencing (Ig-seq) using molecular amplification fingerprinting (MAF). **(A)** In step 1, reverse transcription (RT) is performed to generate first-strand cDNA with a gene-specific (IgM or IgG) primer which includes a unique reverse molecular identifier (RID) and partial Illumina adapter (IA) region. This results in single-molecule labeling of each cDNA with an RID. In step 2, several cycles of multiplex-PCR are performed using a forward primer set with gene-specific regions targeting heavy chain variable (V_H) framework region 1 (FR1), with overhang regions comprised of a forward unique molecular identifier (FID) and partial IA. In step 3, singleplex-PCR is used to extend the partial IAs. The result (Step 4) is the generation of antibody amplicons with FID, RIDs, and full IA ready for Ig-seq and subsequent MAF-based error and bias correction. **(B)** List of oligonucleotides sequences annealing to the V_H FR1 used in multiplex-PCR (Step 2) of the MAF library preparation protocol. The nearest germline IGHV segment(s) likely to be amplified by the respective primer are listed in the rightmost column. **(C)** The estimated melting temperature distribution of the V_H FR1 forward primer set.

Figure 3. Synthetic standards used to validate performance of error and bias correction of Ig-seq data by MAF. **(A)** The number of erroneous CDR3 sequences (at least one a.a. difference from the correct CDR3) per 100,000 reads is plotted against the relative concentration of each standard (from a master stock). Color-coded diamonds correspond to germline IGHV segment family of the respective standard and show the number of erroneous variants in uncorrected (raw) data; gray diamonds indicate the number of variants remaining after MAF error correction. **(B)** The number of erroneous VDJ variants derived from each standard was calculated by finding all variants that carried the correct CDR3 a.a. sequence, but differed by at least one nt across the entire VDJ region. Colored diamonds represent uncorrected data; gray diamonds indicate variants remaining after MAF error correction. **(C)** Sequencing bias introduced by multiplex-PCR using the FR1 primer set was assessed by plotting the measured frequencies of each standard against its relative concentration (from a master stock). Dashed line represents a bias-free ideal scenario ($R^2 = 1$). The left and right plots show observed frequencies before and after MAF bias correction, respectively. **(D)** Phylogenetic trees visualizing the CDR3 a.a. variants present for a selected standard with the CDR3 a.a. sequence *CARGINGERALEW* and IGHV1-8 and IGHJ1 segment usage. Prior to error correction, 39 erroneous CDR3 a.a. variants (branches) and 218 VDJ nt variants (black circles) were observed. Following MAF error correction, only the original correct CDR3 a.a. and two VDJ nt variants remain. The Ig-seq data sets used in A-C consisted of ~300,000 preprocessed full-length antibody reads from each of the synthetic spike-in only samples. IgG1_D1 dataset was used for panel **(D)** (see **Table S3**).

Figure 4. Ig-seq analysis of human naïve ($CD27^-IgM^+$) and memory ($CD27^+IgG^+$) B cells. **(A)** The flow cytometric workflow for isolating $CD27^-IgM^+$ and $CD27^+IgG^+$ B cells from peripheral blood. Boxed-in values indicate the frequency of each sorted subset as a percentage of the total B cell ($CD19^+$) population from each of three donors (1-3 from left to right). **(B)** Experimental and Ig-seq based quantitation of antibody diversity; points represent cDNA molecule counts (using ddPCR) or unique reads (before and after MAF error correction) from cellular replicates (with mean and standard deviation shown) isolated from each donor and B cell subset. Unique read counts were based on the VDJ nt seqeunce. Dashed line represents the number of B cells isolated per cellular replicate (2×10^5 cells). **(C)** Phylogenetic trees illustrating CDR3 a.a. and nt variants present for the

477 selected clonotype with the consensus CDR3 sequence *CARAAGSQYYYMDVW* and IGHV1-8 to
478 IGHJ1 recombination. Prior to error correction, 70 erroneous CDR3 a.a. variants and 249 VDJ nt
479 variants (black circles) were observed. Following MAF error correction, only 6 CDR3 a.a. and 15
480 VDJ nt variants remain. The Ig-seq data sets used in **(B)** are described in **Table S3**; IgG1_D1 was
481 used for the tree in panel **(C)**.
482
483 **Figure 5.** Ig-seq analysis of molecular features highlight global differences between naïve ($CD27^-$
484 IgM^+) and memory ($CD27^+IgG^+$) B cells. **(A)** Clonotype size (calculated as the total number of
485 variants within a clonotype) for naïve IgM (blue lines) and memory IgG (red lines) B cells. Each
486 pair of lines represents a single donor. **(B)** Graph showing the distribution of average SHM
487 frequencies for VDJ nt variants per clonotype. The $CD27^-IgM^+$ B cell repertoires (blue lines) have a
488 median SHM value of zero, whereas only a small fraction of clonotypes (approximately 8%)
489 contain an average of one or more mutations. $CD27^+IgG^+$ repertoires (red lines) have a median of
490 20 to 24 SHM per nt variant within each clonotype. **(C)** CDR3 a.a. length distribution across
491 clonotypes from naïve (blue lines) and memory (red lines) B cell subsets.
492
493 **Figure 6.** Clonotype diversity analysis across cellular replicates of naïve ($CD27^-IgM^+$) and memory
494 ($CD27^+IgG^+$) B cells. **(A)** Venn diagrams show the presence of clonotypes (80% CDR3 a.a.
495 similarity to least one clone in the cluster, same CDR3 a.a. length, same IGHV and IGHJ gene
496 segment usage) across cellular replicates (2 x 10^5 cells each) from each donor B cell subset. **(B)** Bar
497 graph showing the fraction of clonotypes containing at least one variant with at least one nt SHM.
498 Blue and red bars indicate clonotypes identified either in only one aliquot (unique) or in several
499 aliquots (shared), respectively. The p-values represent significance using Fisher's exact test. **(C)**
500 Species accumulation curves for $CD27^+IgG^+$ B cells: the number of newly discovered clonotypes
501 from each additional cellular replicate (black circles) is plotted. Extrapolating the observed overlap
502 provides an estimate for the total number of distinct clonotypes (Chao2 estimator: D1 = 164,268 ±
503 2,365; D2 = 38,034 ± 1,302; D3 = 76,904 ± 1,409) and the approximate amount of cellular
504 replicates needed to discover all clonotypes present in the peripheral blood $CD27^+$ IgG^+ population.
505
506 **Figure 7**. Genetic features of naïve and IgG memory BCR repertoires. **(A)** Block map shows IGHV
507 gene segment usage sorted by family (color-coded) across donors and B cell subset; blocks are
508 normalized to the total number of clonotypes within each group. **(B)** IgG subclass usage in
509 $CD27^+IgG^+$ donor repertoires. IgG_4 sequences (dark blue bars) were virtually absent among three
510 donors; a small fraction of sequences could not be unambiguously mapped based on the sequencing
511 read (gray bars). **(C)** Venn diagrams showing the overlap of clonotypes (80% CDR-H3 amino acid
512 similarity to least one clone in the cluster, same CDR3 a.a. length, same IGHV and IGHJ gene
513 segment usage) between the naïve ($CD27^-IgM^+$) and memory ($CD27^+IgG^+$) BCR heavy chain
514 repertoire in each of three donors. **(D)** Each plot shows the clonal composition of each shared
515 clonotype (from panel **(C)**) in terms of its IgG and IgM intraclonal variants. Red triangles indicate
516 clonotypes found in multiple IgM cellular aliquots; blue triangles show clonotypes which could
517 only be found in one IgM cellular aliquot. The total number of clonotypes found are depicted in the
518 corresponding contingency table. Fisher's exact test was used to quantitatively analyze enrichment
519 of expanded IgM clonotypes in the shared IgM/IgG subset. **(E)** Ternary plots comprised of three
520 axes representing the IgG1, IgG2, and IgG3 isotype subclasses. The relative subclass composition
521 of intraclonal variants per IgG clonotype (each represented by a circle colored according to the
522 number of variants belonging to that clonotype) is depicted by the position of the circle within the

triagonal space. Red circles represent the average subclass composition for all IgG variants of each donor (cf. **Figure 6C**).

METHODS

Preparation of spike-in master stocks

The spike-in standards were ordered from GeneArt (Invitrogen) in the form of plasmids. Each spike-in sequence contained a T7 promoter for *in vitro* transcription. Approximately 1.5 µg of each plasmid was digested with 10 U of EcoRV-HF (New England BioLabs) and purified with DNA-binding magnetic beads (SPRI select, Beckman Coulter). Approximately 1 µg of the digested plasmid was then used for *in vitro* transcription (MEGAscript T7 Transcription Kit, ThermoFisher Scientific). RNA was purified by lithium-chloride precipitation, eluted (TE with 1U/µl RiboLock) and aliquoted. The final concentration was then determined with the TapeStation (Agilent Technologies).
The spike-in RNA obtained this way was reverse transcribed with the Maxima Reverse Transcriptase kit (Thermofisher Scientific). 500 ng mRNA was mixed with 20 pmol of IgM reverse primer and 3 µl dNTP-mix (10 mM each) and was then filled up to 14.5 µl with water. The reaction-mix was incubated at 65 °C for 5 minutes. 4 µl of 5x RT-buffer, 0.5 µl (20 U) RiboLock and 1 µl Maxima reverse transcriptase (200 U) were then added. The resulting reaction mix was incubated for 35 minutes at 55 °C, followed by a termination step at 85 °C for 5 minutes. 2.5 µl of RNase A (Thermofisher Scientific) was added and the mix was again incubated at 60 °C for 30 minutes. The resulting cDNA was purified with SPRI Select magnetic beads and eluted in nuclease-free water. The concentration of each cDNA reaction was determined afterwards with the Fragment Analyzer and pooled according to Table S1. The exact concentration of the pooled spike-ins was determined by ddPCR with dilutions of the pool ranging from 10^{-3} to 10^{-6}. The measured spike-in pool was afterwards diluted to a final storage concentration of 250,000 transcripts per µl.

Transcript quantitation by ddPCR

Quantifiying cDNA and PCR products by ddPCR was conducted for all measurements in the following way: A dilution series with 3 or 4 points was prepared. Droplets were generated with BioRad's droplet generator using 12.25 µl of ddPCR Supermix (BioRad) combined with 10 µl of the diluted sample, 25 pmol of the biological ddPCR probe (SF_21), 25 pmol of the spike-in specific ddPCR probe (TAK_499), 22.5 pmol of the forward (SF_63) and reverse ddPCR primer (TAK_522) and 55 µl of droplet generation oil (BioRad). Droplets were then transferred to a 96-well reaction plate, which was heat sealed with easy pierce foil (VWR International). Then a PCR reaction was performed using the following conditions: 95°C for 10 min; 45 cycles of 94 °C for 30s, 53 °C for 30s, 64 °C for 1 min; 98°C for 10 min; and holding at 4 °C. After the PCR step, every 96-well plate was read using BioRad's droplet reader.

B-cell isolation, sorting, and lysis

569 Peripheral blood leukocyte-enriched fractions ('buffy coats') were received from the Bern
570 (Switzerland) blood donation center after obtaining the proper informed consent from healthy
571 human donors. Blood samples were diluted 1:3 with sterile PBS and overlaid on Ficoll-Paque
572 PLUS (GE Healthcare) using LeukoSep conical centrifuge tubes (Greiner Bio-One). Peripheral
573 blood mononuclear cells (PBMCs) were harvested after separation for 30 minutes at 400 x *g*
574 without braking. Successive centrifugation steps were performed to wash the mononuclear cell
575 fraction and remove residual neutrophils and granulocytes. Total B cells were isolated from PBMCs
576 by negative selection with the EasySep Human B Cell Enrichment Kit (STEMCELL Technologies)
577 according to the manufacturer's instructions. The following fluorescently labeled antibodies were
578 used to stain the enriched B-cell fraction prior to sorting by flow cytometry: anti-CD3-APC/Cy7
579 (clone HIT3a BioLegend # 300318), anti-CD14-APC/Cy7 (clone HCD14 BioLegend #325620),
580 anti-CD16-APC/Cy7 (clone 3G8 BioLegend #302018), anti-CD19-BV785 (clone HIB19,
581 BioLegend #302240), anti-CD20-BV650 (clone 2H7, BioLegend #302336), anti-CD27-V450
582 (clone M-T271, BD Horizon #560448), anti-CD24-BV510 (clone ML5, BioLegend #311126), anti-
583 CD38-PC5 (clone LS198-4-3, Beckman Coulter #A07780), anti-IgD-PEcy7 (clone IA6-2,
584 BioLegend #348210), anti-IgG-Alexa Fluor 647 (Jackson ImmunoResearch #109-606-003) anti-
585 IgA-Alexa Fluor 488 (Jackson ImmunoResearch #109-549-011), anti-IgM-PE (clone SA-DA4,
586 eBioscience #12-9998). Cell sorting was performed on a BD FACS Aria III following the gating
587 strategy depicted in Figure 4A. After sorting, isolated fractions were centrifuged 5 minutes at 300 x
588 *g*, the supernatants were aspirated, and the cell pellets were re-suspended in 1 ml PBS. Recovered
589 cells were hand-counted using a Neubauer hemocytometer, and aliquots containing equal numbers
590 of cells were prepared from the cellular suspension. These aliquots were centrifuged, and the
591 supernatant aspirated. Cell pellets were lysed directly in 200 µl TRI Reagent (Sigma), allowed to
592 dissociate for 5 minutes at room temperature, then frozen on dry ice prior to storage at -80°C.
593
594 **RNA isolation from sorted B-cell populations**
595
596 Immediately prior to use, Phase Lock Gel (PLG) tubes were pelleted at 12000 - 16000 x g in a
597 microcentrifuge for 20 to 30 seconds. Each TRIzol aliquot was then thawed on ice. After thawing
598 and an incubation time of 5 minutes at room temperature, 1 mL of the TRIzol homogenate was
599 transferred to the phase lock tube. 0.2 mL chloroform was added and the tube was shaken
600 vigorously by hand (~15 seconds). After an incubation time of 3 minutes, the phase lock tube was
601 centrifuged at 12'000 x g during 15 minutes at 4 °C. The resulting upper aqueous phase was
602 transferred to a fresh Eppendorf tube, an equal volume of 70% ethanol was added and the solution
603 was purified on a PureLink RNA column according to the manufacturer's instructions (Life
604 Technologies). Finally, RNA was eluted in 25 µl nuclease-free water.
605
606 **Library preparation and NGS on Illumina's MiSeq**
607
608 First-strand cDNA synthesis was carried out using Maxima reverse transcriptase (Life
609 Technologies). The protocol for one reaction is as follows: A 29 µl reaction mix was prepared using
610 up to 2 µg of RNA together with 40 pmol of the respective gene-specific reverse primer (for IgG
611 sequences, 5'-RID-GTTCTGGGAAGTAGTCCTTGACCAG-3' (IgH_10r); for IgM, 5'-RID-
612 ACGAGGGGGAAAAGGGTTGG-3' (CH1_1r)), 2 µl dNTP (10 mM each) and the required
613 amount of nuclease-free water. This mix is incubated for 5 minutes at 65 °C. A master mix of 8 µl
614 5x RT buffer, 1 µl of (20 U) RiboLock and 2 µl of Maxima reverse transcriptase is then prepared

615 and added to the reaction mix. Finally, the mix is incubated for 30 minutes at 50°C and the reaction
616 is terminated by incubating at 85°C for 5 minutes.
617 The obtained cDNA is then cleaned with a left-sided SPRI-Select bead clean up (0.8x) according to
618 the manufacturer's instructions (Beckman coulter) and subsequently measured by ddPCR.
619 Up to 135,000 cDNA transcripts were then pooled together with 12,500 spike-in transcripts.
620 Multiplex-PCR was performed using an equimolar pool of the forward primer mix (Figure 2B) and
621 the reverse primer (TAK_423) targeting the overhang introduced during cDNA synthesis. Due to
622 low cDNA yields, the first PCR was carried out for 20 cycles and the following cycling protocol: 2
623 min at 95 °C; 20 cycles of 98°C for 20s; 60°C for 50; 72°C for 1min; 72°C and then holding at 4°C.
624 PCR reactions were prepared using 15 µl of Kapa HiFi HotStart ReadyMix (KAPA Biosystems), 50
625 pmol of the forward mix and the reverse primer and 9 µl of the cDNA mix. After the first PCR, we
626 again performed a left-sided bead clean-up (0.8x) and measured the PCR product concentration
627 using ddPCR. We use 800,000 transcripts from PCR 1 as input into the adapter extension PCR. For
628 this PCR 25 µl of Kapa HiFI Hotstart ReadyMix was combined with 25 pmol of the forward primer
629 (TAK_424) and 25 pmol of the index primer (TAK_531) as well as the diluted PCR product.
630 Finally, the reaction volume was adjusted to 50 µl by the addition of nuclease-free water.
631 Thermocycling was performed as follows: 95°C for 5 min; 23 cycles of 98°C for 20 s, 65°C for 15
632 s, 72°C for 15 s; 72°C for 5 min; and 4°C indefinitely. Following second-step adapter extension
633 PCR, reactions were cleaned using a double-sided SPRIselect bead cleanup process (0.5x to 0.8x),
634 with an additional ethanol wash and elution in TE buffer.
635 Libraries were then quantified by capillary electrophoresis (Fragment analyzer, Agilent). After
636 quantitation, libraries were pooled accordingly and sequenced on a MiSeq System (Illumina) with
637 the paired-end 2x300bp kit.
638
639 **Bioinformatic pipeline**
640
641 Paired-end fastq files were merged using PandaSeq (40). Afterwards, sequences were filtered for
642 quality and length using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). After the
643 quality trim, sequences were processed with a custom Python script that performed error correction
644 by consensus building on our sequences and RIDs. In order to utilize as many sequencing reads as
645 possible, we required UIDs to have at least 3 reads, but did not remove sequences that only had one
646 UID group mapping to them. VDJ annotation and frequency calculation was then performed by our
647 in-house aligner (18) which was updated with the human reference database downloaded from
648 IMGT. The complete error-correction and alignment pipeline is available under
649 https://gitlab.ethz.ch/reddy/MAF.
650
651 **Statistical analysis**
652
653 All statistical and computational analyses following the alignment step were performed in R.
654 Details about specific tests that were used can be found in the results section and in the figure
655 legends. Scripts are available upon request.
656
657 **Data availability**
658
659 In adherence to the data sharing recommendations of the AIRR community our data is publically
660 available in the following repositories: BioProject, BioSample, SRA and GenBank and can be
661 accessed with the accession number PRJNA430091 (BioProject). The exact data processing steps,

662 including software tools and version numbers can be found on zonodo.org under the following doi:
663 10.5281/zenodo.1201416.
664
665 Likewise, the designed spike-in sequences are also stored on GenBank (Accession number
666 MG785894-MG785978).
667

668 **Author contributions**

669 J.M.L., S.F., E.T., and S.T.R. designed experiments. V.C. performed B-cell enrichment, sorting,
670 and mRNA extraction. M.I. and S.F. prepared IgH libraries. J.M.L. designed primer sequences.
671 J.M.L. and A.Z. designed antibody spike-ins. A.Z., S.M., and M.I. conducted preliminary
672 experiments. E.M. analysed preliminary data. S.F. was responsible for the bioinformatics pipeline.
673 J.M.L. and S.F. analysed the final data and prepared figures. J.M.L., S.F., E.T., and S.T.R. wrote the
674 paper. All authors provided scientific guidance.

675

676 **Acknowledgments**

677

678 We would like to acknowledge the Genomics Facility Basel of ETH Zurich for Illumina sequencing
679 support, in particular E. Burcklen, K. Eschbach and C. Beisel. We also want to thank H.
680 Ruscheweyh for bioinformatic code support.
681

682 **REFERENCES**

683

684 1. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin
685 repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010)
686 116(7):1070-8. doi: 10.1182/blood-2010-03-275859. PubMed PMID: 20457872; PubMed Central PMCID:
687 PMCPMC2938129.
688 2. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody
689 heavy-chain repertoires in humans. *PLoS ONE* (2011) 6(8):e22365. doi: 10.1371/journal.pone.0022365. PubMed PMID:
690 21829618; PubMed Central PMCID: PMCPMC3150326.
691 3. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and
692 analysis of the human paired heavy- and light-chain antibody repertoire. *Nature medicine* (2015) 21(1):86-91. doi:
693 10.1038/nm.3743. PubMed PMID: 25501908.
694 4. Robinson WH. Sequencing the functional antibody repertoire--diagnostic and therapeutic discovery. *Nat Rev*
695 *Rheumatol* (2015) 11(3):171-82. doi: 10.1038/nrrheum.2014.220. PubMed PMID: 25536486; PubMed Central PMCID:
696 PMCPMC4382308.
697 5. Williams LD, Ofek G, Schätzle S, McDaniel JR, Lu X, Nicely NI, et al. Potent and broad HIV-neutralizing
698 antibodies in memory B cells and plasma. *Science immunology* (2017) 2(7):eaal2200. doi: 10.1126/sciimmunol.aal2200.
699 PubMed PMID: 28783671.
700 6. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway
701 for potent V1V2-directed HIV-neutralizing antibodies. *Nature* (2014) 509(7498):55-62. doi: 10.1038/nature13036.
702 PubMed PMID: 24590074; PubMed Central PMCID: PMCPMC4395007.
703 7. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1-neutralizing
704 antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National*
705 *Academy of Sciences* (2013) 110(16):6470-5. doi: 10.1073/pnas.1219320110. PubMed PMID: 23536288; PubMed
706 Central PMCID: PMCPMC3631616.
707 8. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of
708 the constituent human serum antibodies elicited by vaccination. *Proceedings of the National Academy of Sciences*
709 (2014) 111(6):2259-64. doi: 10.1073/pnas.1317793111. PubMed PMID: 24469811; PubMed Central PMCID:
710 PMCPMC3926051.
711 9. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He X-S, et al. Lineage structure of the human antibody
712 repertoire in response to influenza vaccination. *Science translational medicine* (2013) 5(171):171ra19-ra19. doi:
713 10.1126/scitranslmed.3004794. PubMed PMID: 23390249; PubMed Central PMCID: PMCPMC3699344.

714 10. Roskin KM, Simchoni N, Liu Y, Lee J-Y, Seo K, Hoh RA, et al. IgH sequences in common variable immune
715 deficiency reveal altered B cell development and selection. *Science translational medicine* (2015) 7(302):302ra135-
716 302ra135. doi: 10.1126/scitranslmed.aab1216. PubMed PMID: 26311730; PubMed Central PMCID: PMCPMC4584259.
717 11. Palanichamy A, Apeltsin L, Kuo TC, Sirota M, Wang S, Pitts SJ, et al. Immunoglobulin class-switched B cells
718 form an active immune axis between CNS and periphery in multiple sclerosis. *Science translational medicine* (2014)
719 6(248):248ra106-248ra106. doi: 10.1126/scitranslmed.3008930. PubMed PMID: 25100740; PubMed Central PMCID:
720 PMCPMC4176763.
721 12. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, et al. An atlas of B-cell clonal distribution
722 in the human body. *Nature Biotechnology* (2017) 35(9):879-84. doi: 10.1038/nbt.3942. PubMed PMID: 28829438;
723 PubMed Central PMCID: PMCPMC5679700.
724 13. Lee YN, Frugoni F, Dobbs K, Tirosh I, Du L, Ververs FA, et al. Characterization of T and B cell repertoire
725 diversity in patients with RAG deficiency. *Sci Immunol* (2016) 1(6). doi: 10.1126/sciimmunol.aah6109. PubMed PMID:
726 28783691; PubMed Central PMCID: PMCPMC5586490.
727 14. Friedensohn S, Khan TA, Reddy ST. Advanced Methodologies in High-Throughput Sequencing of Immune
728 Repertoires. *Trends in biotechnology* (2017) 35(3):203-14. doi: 10.1016/j.tibtech.2016.09.010. PubMed PMID: 28341036.
729 15. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of
730 high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* (2014) 32(2):158-68. doi:
731 10.1038/nbt.2782. PubMed PMID: 24441474; PubMed Central PMCID: PMCPMC4113560.
732 16. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq
733 experiments. *Genome Research* (2011) 21(9):1543-51. doi: 10.1101/gr.121095.111. PubMed PMID: 21816910; PubMed
734 Central PMCID: PMCPMC3166838.
735 17. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-
736 free profiling of immune repertoires. *Nature Methods* (2014) 11(6):653-5. doi: 10.1038/nmeth.2960. PubMed PMID:
737 24793455.
738 18. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and
739 predictive antibody repertoire profiling by molecular amplification fingerprinting. *Science advances* (2016) 2(3):e1501371.
740 doi: 10.1126/sciadv.1501371. PubMed PMID: 26998518; PubMed Central PMCID: PMCPMC4795664.
741 19. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with
742 massively parallel sequencing. *Proceedings of the National Academy of Sciences* (2011) 108(23):9530-5. doi:
743 10.1073/pnas.1105422108. PubMed PMID: 21586637; PubMed Central PMCID: PMCPMC3111315.
744 20. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and
745 amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences* (2012)
746 109(4):1347-52. doi: 10.1073/pnas.1118018109. PubMed PMID: 22232676; PubMed Central PMCID:
747 PMCPMC3268301.
748 21. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of
749 molecules using unique molecular identifiers. *Nature Methods* (2011) 9(1):72-4. doi: 10.1038/nmeth.1778. PubMed
750 PMID: 22101854.
751 22. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using
752 antibody repertoire sequencing. *Proceedings of the National Academy of Sciences* (2013) 110(33):13463-8. doi:
753 10.1073/pnas.1312146110. PubMed PMID: 23898164; PubMed Central PMCID: PMCPMC3746854.
754 23. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length
755 immunoglobulin profiling with unique molecular barcoding. *Nature Protocols* (2016) 11(9):1599-616. doi:
756 10.1038/nprot.2016.093. PubMed PMID: 27490633.
757 24. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. Highly Accurate Sequencing of Full-Length
758 Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier-Guided Amplicon Assembly. *The Journal of*
759 *Immunology* (2016) 196(6):2902-7. doi: 10.4049/jimmunol.1502563. PubMed PMID: 26856699.
760 25. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international
761 ImMunoGeneTics information system® 25 years on. *Nucleic acids research* (2015) 43(Database issue):D413-22. doi:
762 10.1093/nar/gku1056. PubMed PMID: 25378316; PubMed Central PMCID: PMCPMC4383898.
763 26. Wang C, Liu Y, Xu LT, Jackson KJL, Roskin KM, Pham TD, et al. Effects of aging, cytomegalovirus infection,
764 and EBV infection on human B cell repertoires. *The Journal of Immunology* (2014) 192(2):603-11. doi:
765 10.4049/jimmunol.1301384. PubMed PMID: 24337376; PubMed Central PMCID: PMCPMC3947124.
766 27. Menzel U, Greiff V, Khan TA, Haessler U, Hellmann I, Friedensohn S, et al. Comprehensive evaluation and
767 optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS ONE* (2014)
768 9(5):e96727. doi: 10.1371/journal.pone.0096727. PubMed PMID: 24809667; PubMed Central PMCID:
769 PMCPMC4014543.
770 28. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA, et al. Quantitative assessment of the
771 robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunology*
772 (2014) 15(1):40. doi: 10.1186/s12865-014-0040-5. PubMed PMID: 25318652; PubMed Central PMCID:
773 PMCPMC4233042.
774 29. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution
775 antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences* (2014)
776 111(13):4928-33. doi: 10.1073/pnas.1323862111. PubMed PMID: 24639495; PubMed Central PMCID:
777 PMCPMC3977259.
778 30. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires.
779 *Trends in immunology* (2015) 36(11):738-49. doi: 10.1016/j.it.2015.09.006. PubMed PMID: 26508293.

31. Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, Chazdon RL, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* (2012) 5(1):3-21. doi: 10.1093/jpe/rtr044.
32. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature Biotechnology* (2010) 28(9):965-9. doi: 10.1038/nbt.1673. PubMed PMID: 20802495.
33. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung M-W, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nature communications* (2013) 4:2680. doi: 10.1038/ncomms3680. PubMed PMID: 24157944.
34. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nature medicine* (2016) 22(12):1456-64. doi: 10.1038/nm.4224. PubMed PMID: 27820605; PubMed Central PMCID: PMCPMC5301914.
35. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. *Science (New York, NY)* (2015) 348(6239):aaa0698. doi: 10.1126/science.aaa0698. PubMed PMID: 26045439; PubMed Central PMCID: PMCPMC4844011.
36. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLoS ONE* (2016) 11(8):e0160853. doi: 10.1371/journal.pone.0160853. PubMed PMID: 27513338; PubMed Central PMCID: PMCPMC4981401.
37. Vidarsson G, Dekkers G, Rispens T. IgG subclasses and allotypes: from structure to effector functions. *Frontiers in immunology* (2014) 5(16):520. doi: 10.3389/fimmu.2014.00520. PubMed PMID: 25368619; PubMed Central PMCID: PMCPMC4202688.
38. Heeringa JJ, Karim AF, van Laar JAM, Verdijk RM, Paridaens D, van Hagen PM, et al. Expansion of blood IgG4(+) B, TH2, and regulatory T cells in patients with IgG4-related disease. *J Allergy Clin Immunol* (2017). doi: 10.1016/j.jaci.2017.07.024. PubMed PMID: 28830675.
39. Torres M, Casadevall A. The immunoglobulin constant region contributes to affinity and specificity. *Trends Immunol* (2008) 29(2):91-7. doi: 10.1016/j.it.2007.11.004. PubMed PMID: 18191616.
40. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC bioinformatics* (2012) 13(1):31. doi: 10.1186/1471-2105-13-31. PubMed PMID: 22333067; PubMed Central PMCID: PMCPMC3471323.

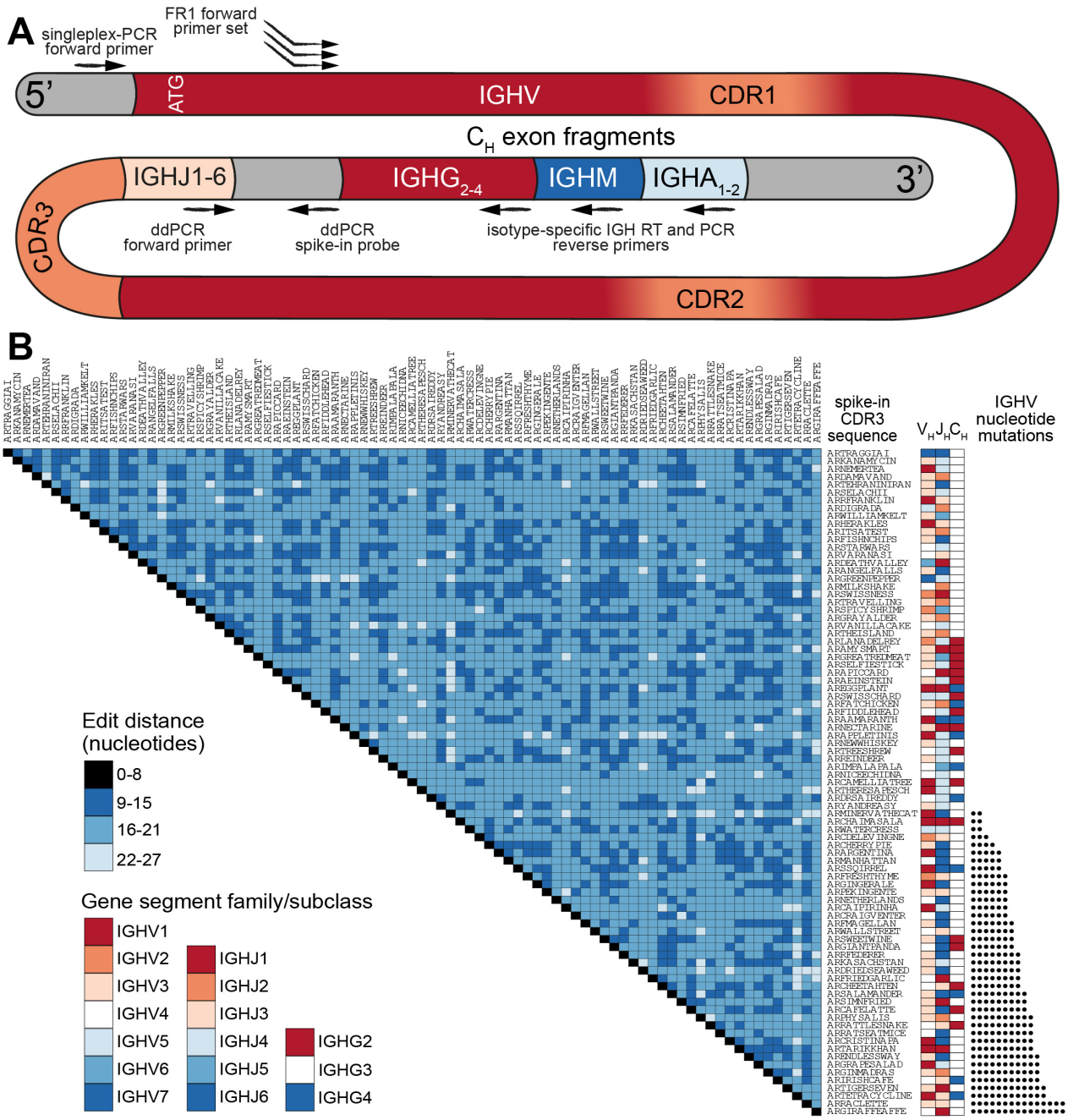
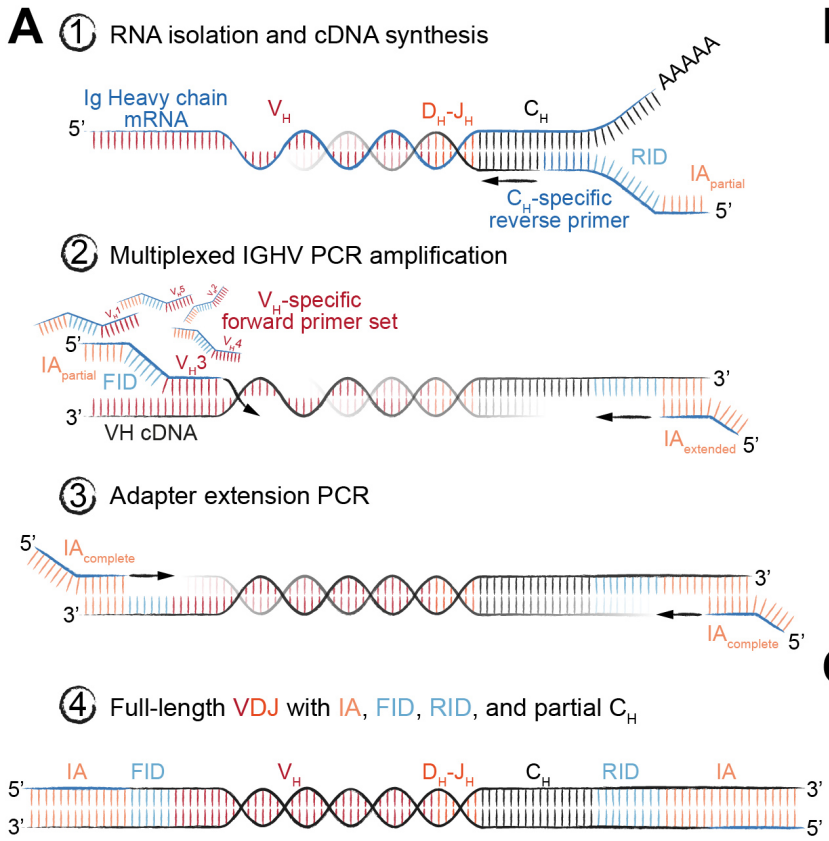


Figure 1. A comprehensive set of human synthetic spike-in standards for Ig-seq. **(A)** Schematic showing the prototypical spike-in with the following regions (5' to 3'): a conserved non-coding region, ATG start codon, IGHV region with FR1 specific for multiplex-PCR, IGHJ regions, non-coding synthetic sequence identifier (specific for ddPCR probes), and downstream heavy chain constant domain sequences (IGHG, IGHM, IGHA) containing primer binding sites used for cDNA synthesis. Each spike-in contains a complete VDJ open reading frame, including nucleotides upstream of the ATG start codon, and downstream constant domain sequences containing primer binding sites used for sample cDNA synthesis. **(B)** Pairwise comparisons based on a.a. Levenshtein edit distance of all 85 standard CDR3 sequences. The germline IGHV and IGHJ segment family usage and IgG subclasses are denoted. 39 spike-ins contain rationally designed nt SHM (black circles) across the IGHV regions.



B

primer name and aligned sequence		IGHV segment coverage
VH1a	GCAGCTGGTGCAGTCTGGGG	1-2,1-3,1-8,1-45, 1-46,1-69
VH1b	TCAGCTGGTGCAGTCTGGAG	1-18
VH1c	CCAGCTTGTGCAGTCTGGGG	1-24
VH1d	GCAGCTGGTGCAGTCTGGGC	1-58
VH2a	CCCAGGTCACCTTGAAGGAGTCTG	2-26, 2-70
VH2b	CCCAGATCACCTTGAAGGAGTCTG	2-5
VH3a	TGTGAGGTGCAGCTGGTGGAGTC	3-7, 3-13, 3-15, 3-20, 3-21, 3-23, 3-48, 3-53, 3-64, 3-66, 3-72, 3-73, 3-74
VH3b	TGTGAAGTGCAGCTGGTGGAGTC	3-9, 3-43
VH3c	TGTCAGGTGCAGCTGGTGGAGTC	3-11, 3-30, 3-33
VH3d	TGTGAGGTGCAGCTGGTGGAGAC	3-53*2
VH4a	TGCAGCTGCAGGAGTCGGG	4-4, 4-28, 4-30-4, 4-31, 4-38, 4-39, 4-49, 4-59, 4-61
VH4b	TACAGCTGCAGGAGTCGGG	4-28*7
VH4c	TGCAGCTGCAGGAGTCCGG	4-30-2
VH4d	TGCAGCTACAGCAGTGGGG	4-34
VH5	GCAGCTGGTGCAGTCTGGAG	5-10-1, 5-51
VH6	TACAGCTGCAGCAGTCAGG	6-1
VH7	GCAGCTGGTGCAATCTGGGT	7-4-1, 7-81

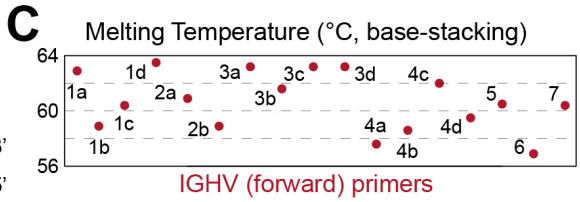


Figure 2. Library preparation of immunoglobulin (Ig) heavy chain genes for high-throughput sequencing (Ig-seq) using molecular amplification fingerprinting (MAF). **(A)** In step 1, reverse transcription (RT) is performed to generate first-strand cDNA with a gene-specific (IgM or IgG) primer which includes a unique reverse molecular identifier (RID) and partial Illumina adapter (IA) region. This results in single-molecule labeling of each cDNA with an RID. In step 2, several cycles of multiplex-PCR are performed using a forward primer set with gene-specific regions targeting heavy chain variable (V_H) framework region 1 (FR1), with overhang regions comprised of a forward unique molecular identifier (FID) and partial IA. In step 3, singleplex-PCR is used to extend the partial IAs. The result (Step 4) is the generation of antibody amplicons with FID, RIDs, and full IA ready for Ig-seq and subsequent MAF-based error and bias correction. **(B)** List of oligonucleotides sequences annealing to the V_H FR1 used in multiplex-PCR (Step 2) of the MAF library preparation protocol. The nearest germline IGHV segment(s) likely to be amplified by the respective primer are listed in the rightmost column. **(C)** The estimated melting temperature distribution of the V_H FR1 forward primer set.

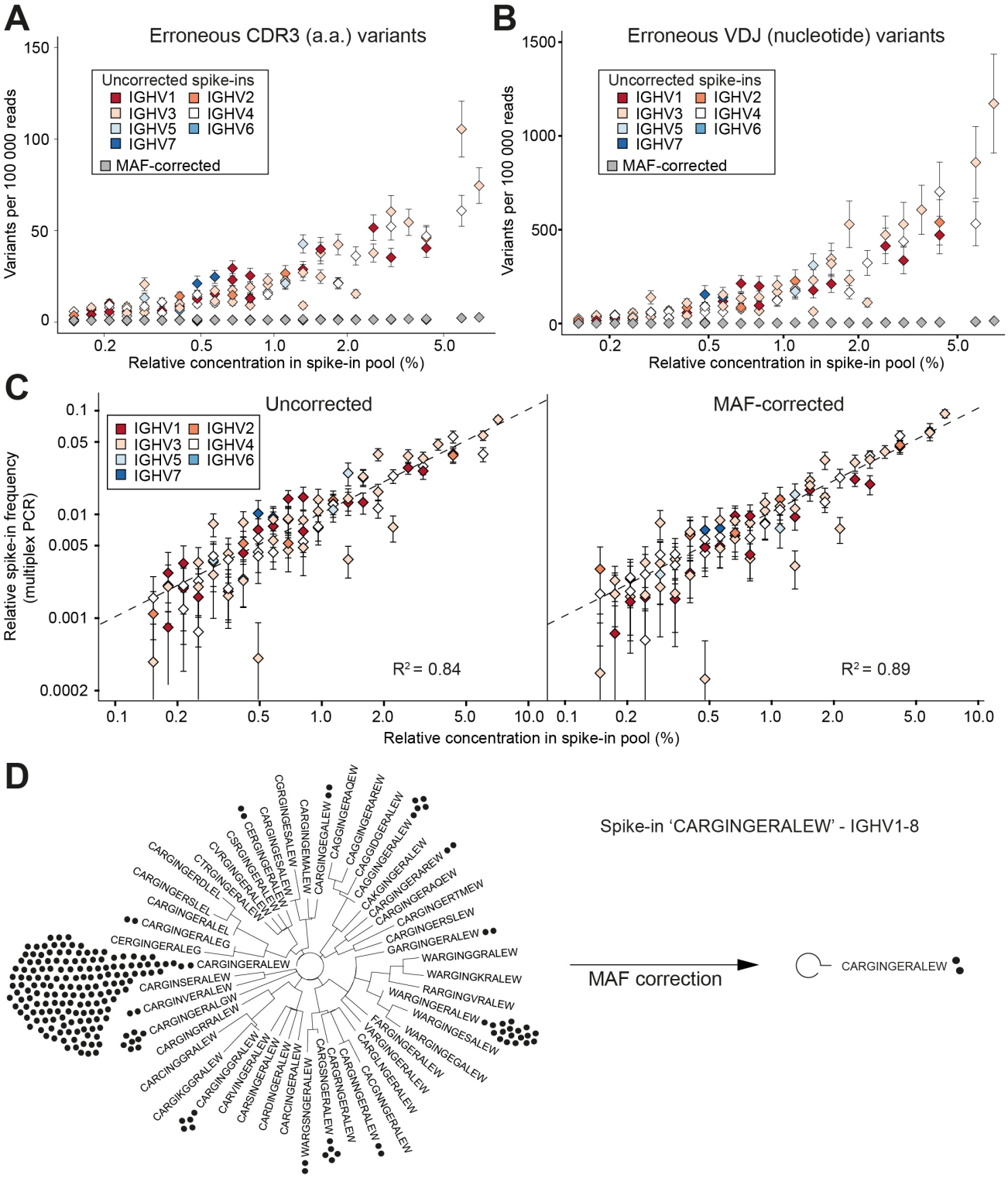


Figure 3. Synthetic standards used to validate performance of error and bias correction of Ig-seq data by MAF. **(A)** The number of erroneous CDR3 sequences (at least one a.a. difference from the correct CDR3) per 100,000 reads is plotted against the relative concentration of each standard (from a master stock). Color-coded diamonds correspond to germline IGHV segment family of the respective standard and show the number of erroneous variants in uncorrected (raw) data; gray diamonds indicate the number of variants remaining after MAF error correction. **(B)** The number of erroneous VDJ variants derived from each standard was calculated by finding all variants that carried the correct CDR3 a.a. sequence, but differed by at least one nt across the entire VDJ region. Colored diamonds represent uncorrected data; gray diamonds indicate variants remaining after MAF error correction. **(C)** Sequencing bias introduced by multiplex-PCR using the FR1 primer set was assessed by plotting the measured frequencies of each standard against its relative concentration (from a master stock). Dashed line represents a bias-free ideal scenario ($R^2 = 1$). The left and right plots show observed frequencies before and after MAF bias correction, respectively. **(D)** Phylogenetic trees visualizing the CDR3 a.a. variants present for a selected standard with the CDR3 a.a. sequence *CARGINGERALEW* and IGHV1-8 and IGHJ1 segment usage. Prior to error correction, 39 erroneous CDR3 a.a. variants (branches) and 218 VDJ nt variants (black circles) were observed. Following MAF error correction, only the original correct CDR3 a.a. and two VDJ nt variants remain. The Ig-seq data sets used in A-C consisted of ~300,000 preprocessed full-length antibody reads from each of the synthetic spike-in only samples. IgG1_D1 dataset was used for panel **(D)** (see **Table S3**).

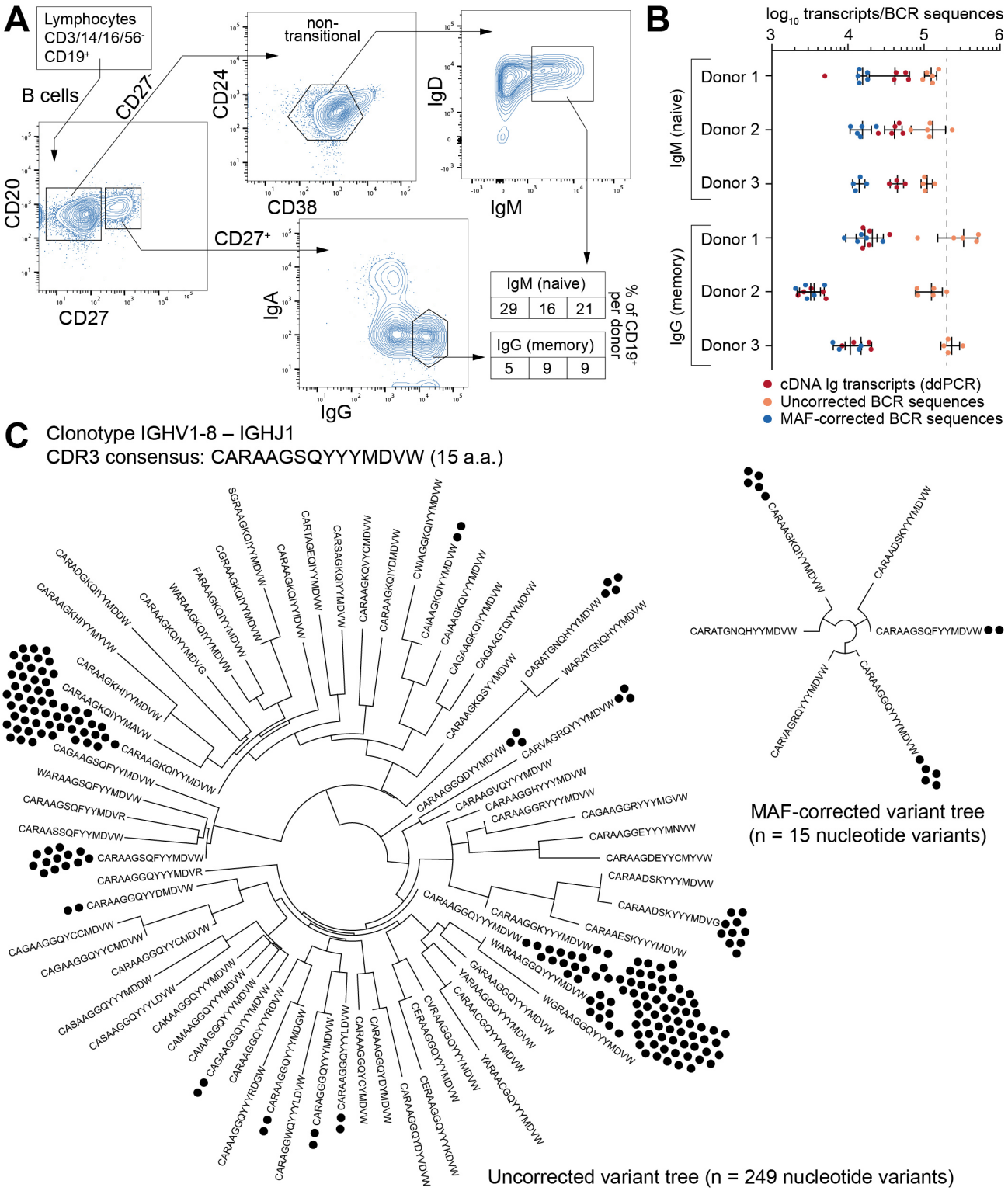


Figure 4. Ig-seq analysis of human naïve (CD27-IgM+) and memory (CD27+IgG+) B cells. **(A)** The flow cytometric workflow for isolating CD27-IgM+ and CD27+IgG+ B cells from peripheral blood. Boxed-in values indicate the frequency of each sorted subset as a percentage of the total B cell (CD19+) population from each of three donors (1-3 from left to right). **(B)** Experimental and Ig-seq based quantitation of antibody diversity; points represent cDNA molecule counts (using ddPCR) or unique reads (before and after MAF error correction) from cellular replicates (with mean and standard deviation shown) isolated from each donor and B cell subset. Unique read counts were based on the VDJ nt seqeunce. Dashed line represents the number of B cells isolated per cellular replicate (2 x 10^5 cells). **(C)** Phylogenetic trees illustrating CDR3 a.a. and nt variants present for the selected clonotype with the consensus CDR3 sequence *CARAAGSQYYYMDVW* and IGHV1-8 to IGHJ1 recombination. Prior to error correction, 70 erroneous CDR3 a.a. variants and 249 VDJ nt variants (black circles) were observed. Following MAF error correction, only 6 CDR3 a.a. and 15 VDJ nt variants remain. The Ig-seq data sets used in **(B)** are described in **Table S3**; IgG1_D1 was used for the tree in panel **(C)**.

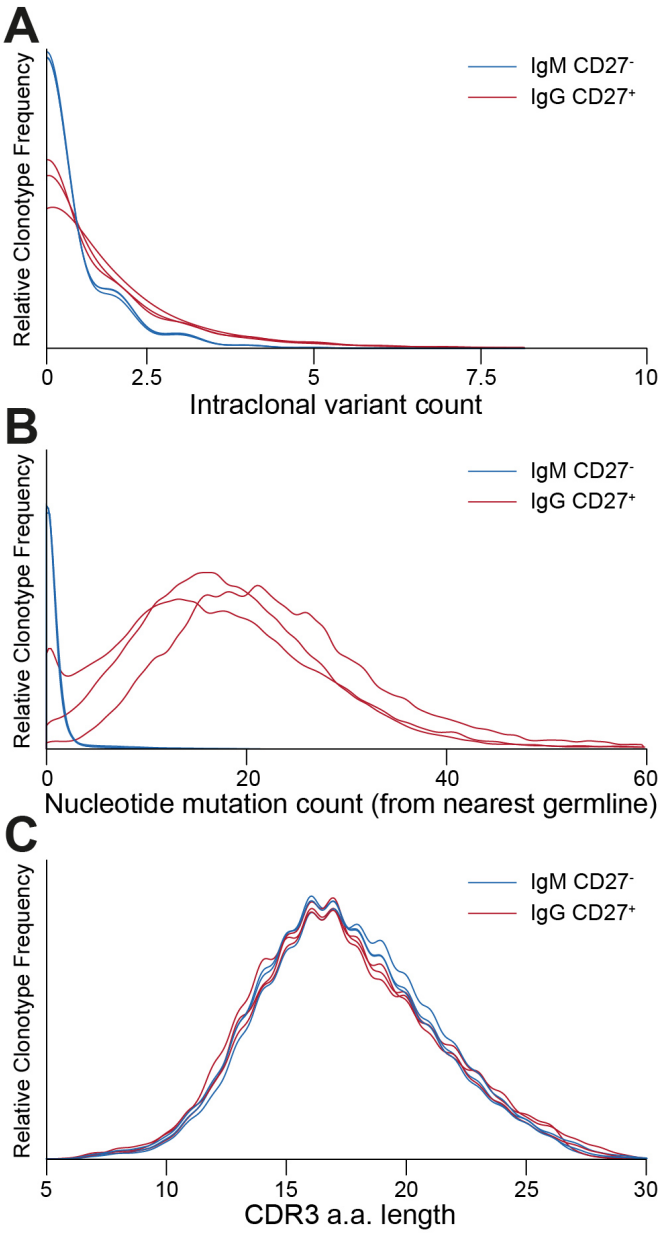


Figure 5. Ig-seq analysis of molecular features highlight global differences between naïve ($CD27^{-}IgM^{+}$) and memory ($CD27^{+}IgG^{+}$) B cells. **(A)** Clonotype size (calculated as the total number of variants within a clonotype) for naïve IgM (blue lines) and memory IgG (red lines) B cells. Each pair of lines represents a single donor. **(B)** Graph showing the distribution of average SHM frequencies for VDJ nt variants per clonotype. The $CD27^{-}IgM^{+}$ B cell repertoires (blue lines) have a median SHM value of zero, whereas only a small fraction of clonotypes (approximately 8%) contain an average of one or more mutations. $CD27^{+}IgG^{+}$ repertoires (red lines) have a median of 20 to 24 SHM per nt variant within each clonotype. **(C)** CDR3 a.a. length distribution across clonotypes from naïve (blue lines) and memory (red lines) B cell subsets.

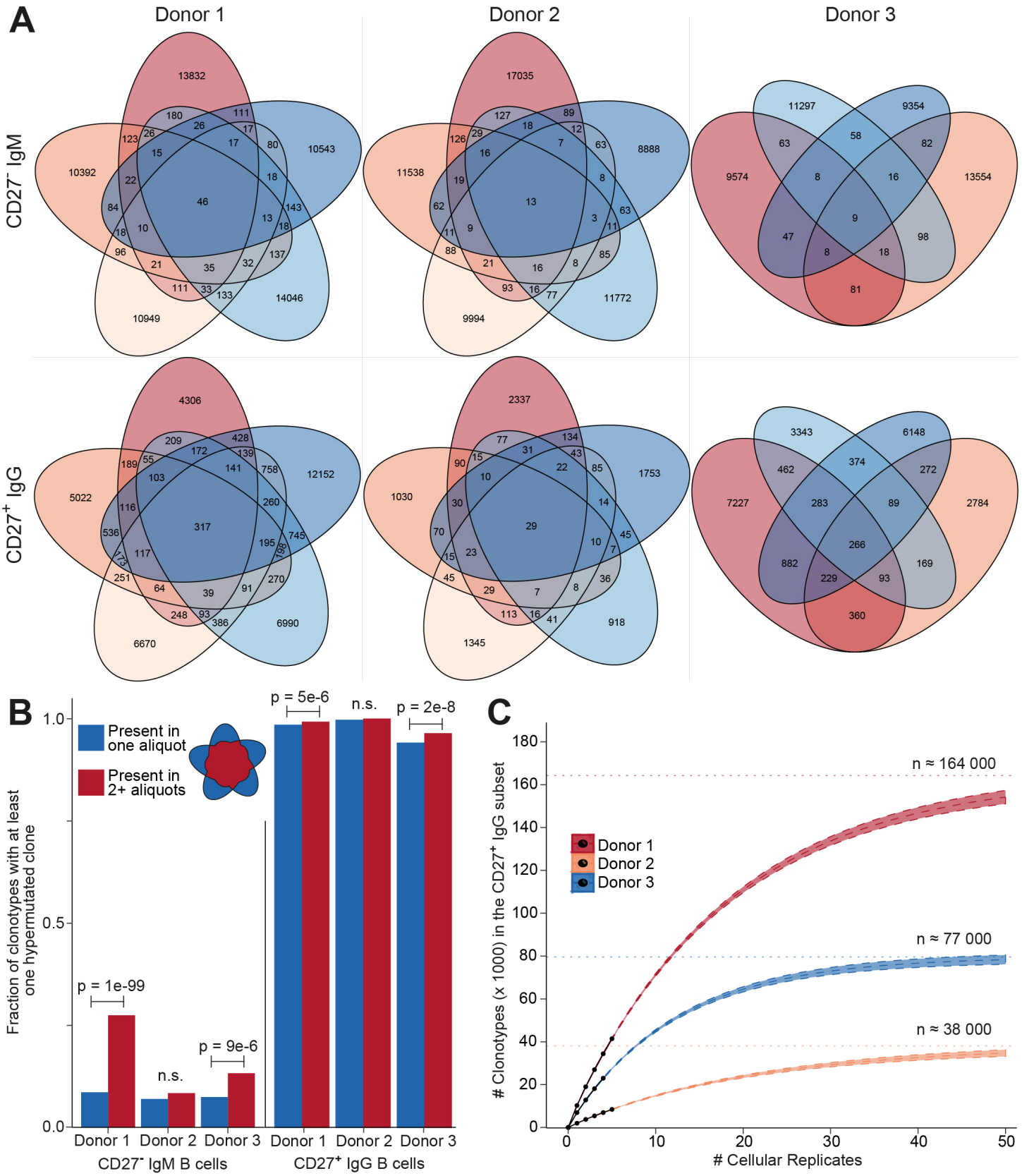


Figure 6. Clonotype diversity analysis across cellular replicates of naïve ($CD27^-IgM^+$) and memory ($CD27^+IgG^+$) B cells. **(A)** Venn diagrams show the presence of clonotypes (80% CDR3 a.a. similarity to least one clone in the cluster, same CDR3 a.a. length, same IGHV and IGHJ gene segment usage) across cellular replicates (2 x 10^5 cells each) from each donor B cell subset. **(B)** Bar graph showing the fraction of clonotypes containing at least one variant with at least one nt SHM. Blue and red bars indicate clonotypes identified either in only one aliquot (unique) or in several aliquots (shared), respectively. The p-values represent significance using Fisher's exact test. **(C)** Species accumulation curves for $CD27^+IgG^+$ B cells: the number of newly discovered clonotypes from each additional cellular replicate (black circles) is plotted. Extrapolating the observed overlap provides an estimate for the total number of distinct clonotypes (Chao2 estimator: D1 = 164,268 ± 2,365; D2 = 38,034 ± 1,302; D3 = 76,904 ± 1,409) and the approximate amount of cellular replicates needed to discover all clonotypes present in the peripheral blood $CD27^+$ IgG^+ population.

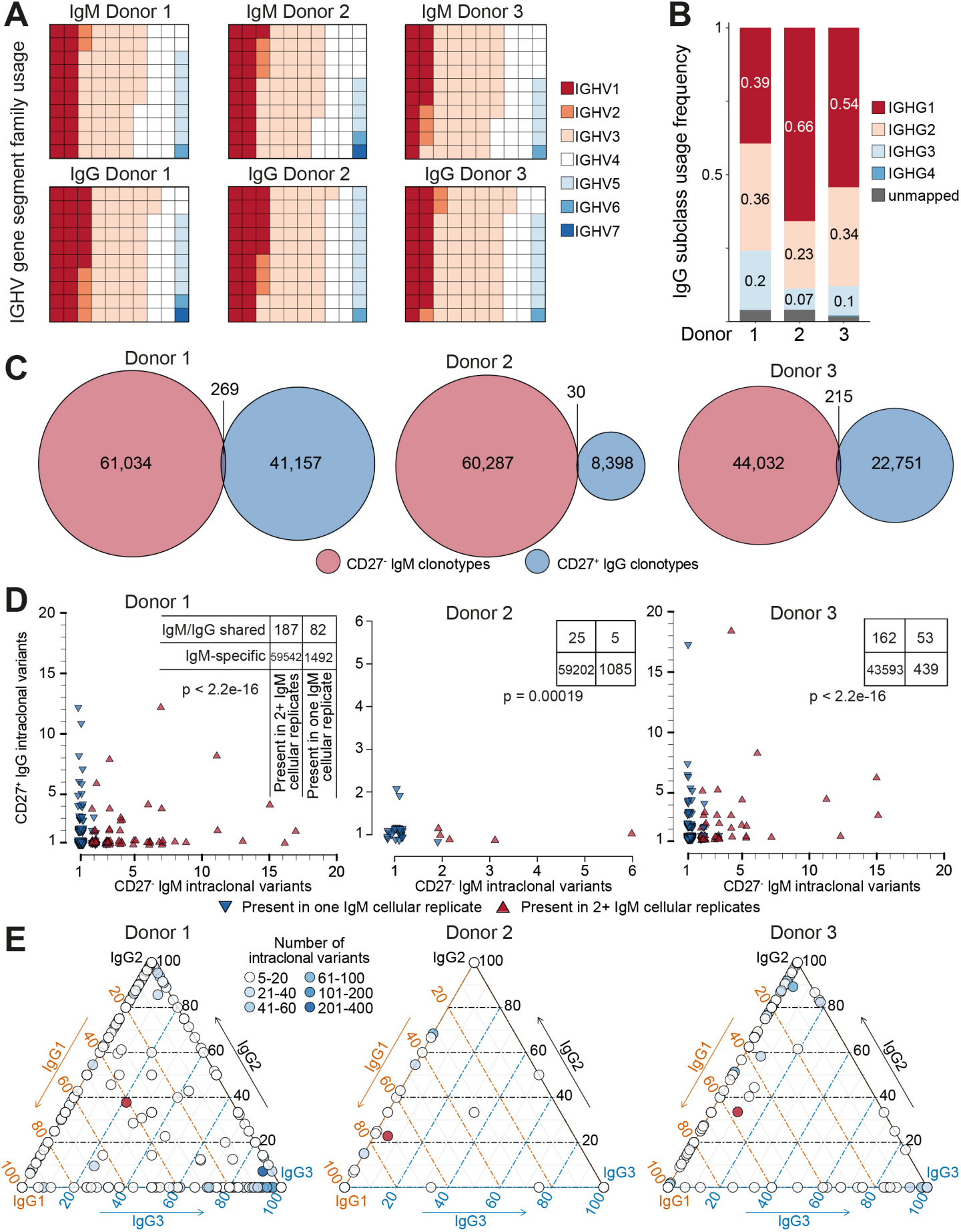


Figure 7. Genetic features of naïve and IgG memory BCR repertoires. **(A)** Block map shows IGHV gene segment usage sorted by family (color-coded) across donors and B cell subset; blocks are normalized to the total number of clonotypes within each group. **(B)** IgG subclass usage in CD27⁺IgG⁺ donor repertoires. IgG_4 sequences (dark blue bars) were virtually absent among three donors; a small fraction of sequences could not be unambiguously mapped based on the sequencing read (gray bars). **(C)** Venn diagrams showing the overlap of clonotypes (80% CDR-H3 amino acid similarity to least one clone in the cluster, same CDR3 a.a. length, same IGHV and IGHJ gene segment usage) between the naïve (CD27⁻IgM⁺) and memory (CD27⁺IgG⁺) BCR heavy chain repertoire in each of three donors. **(D)** Each plot shows the clonal composition of each shared clonotype (from panel **(C)**) in terms of its IgG and IgM intraclonal variants. Red triangles indicate clonotypes found in multiple IgM cellular aliquots; blue triangles show clonotypes which could only be found in one IgM cellular aliquot. The total number of clonotypes found are depicted in the corresponding contingency table. Fisher's exact test was used to quantitatively analyze enrichment of expanded IgM clonotypes in the shared IgM/IgG subset. **(E)** Ternary plots comprised of three axes representing the IgG1, IgG2, and IgG3 isotype subclasses. The relative subclass composition of