FLAIRR-seq: A novel method for single molecule resolution of near full-length immunoglobulin heavy chain repertoires

Authors:

Easton E. Ford[1,2*], David Tieri[2], Oscar Rodriguez[2], Nancy Francoeur[3], Juan Soto[3], Justin Kos[2], William Gibson[2], Catherine A. Silver[2], Gintaras Deikus[3], Elizabeth Hudson[2], Cassandra R. Woolley[1], Noam Beckmann[3], Alexander Charney[3], Thomas C. Mitchell[1], Robert P. Sebra[3], Corey T. Watson[2*], Melissa L. Smith[2*]

Affiliations:

* denotes equal author contribution to the manuscript

[1]Department of Microbiology and Immunology University of Louisville School of Medicine, Louisville, Kentucky, USA

[2]Department of Biochemistry and Molecular Genetics University of Louisville School of Medicine, Louisville, Kentucky, USA

[3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, New York, USA

Corresponding Author:

Melissa L. Smith, PhD

ml.smith@louisville.edu

Abstract:

Current Adaptive Immune Receptor Repertoire Sequencing (AIRR-seq) strategies resolve immunoglobulin (IG) variable region transcripts with limited resolution of the constant region. Here we present a novel near full-length AIRR-seq (FLAIRR-Seq) method that utilizes targeted heavy chain amplification by 5' rapid amplification of cDNA ends (RACE), combined with single molecule, real-time (SMRT) sequencing to generate highly accurate (>Q40, 99.99%), immunoglobulin heavy chain transcripts. FLAIRR-seq was benchmarked by comparing V, D, and J gene usage, complementarity-determining region 3 (CDR3) length, and the extent of total variable gene somatic hypermutation to matched datasets generated with standard 5' RACE AIRR-seq and full-length isoform sequencing methods. Together these data demonstrate robust, unbiased FLAIRR-seq performance using RNA samples derived from peripheral blood mononuclear cells, purified B cells, and whole blood, which recapitulated results generated by commonly used methods, while resolving novel constant region features. FLAIRR-seq data provides, for the first time, simultaneous, single-molecule characterization of variable and constant region genes and alleles, allele-resolved subisotype definition, and identification of class-switch recombination within a clonal lineage. FLAIRR-seq, in conjunction with IGHC genotyping, of the IgM- and IgG repertoires from 10 individuals resulted in the identification of 32 unique IGHC alleles; 28/32 (87%) that were not represented by sequences curated in the IMmunoGeneTics Information System (IMGT) database. Together, these data demonstrate the capabilities of FLAIRR-seq to characterize IGHV, D, J and C gene diversity for the most comprehensive view of B cell repertoires to date.

53 **Introduction**

54 Antibodies or immunoglobulins (IG) are the primary effectors of humoral immunity and are
55 found as both membrane-bound receptors on B cells and circulating, secreted proteins (1).
56 Membrane-bound IG (also known as the B cell receptor (BCR)) and secreted IG act to recognize
57 and bind specific antigen. All IG are composed of a heavy chain and a light chain that are post-
58 translationally associated to complete the correctly folded IG. The heavy chain is comprised of
59 two distinct domains: (i) the variable domain (Fab), which allows for antigen binding, and (ii) the
60 constant domain (Fc), which modulates downstream effector functions (1, 2). The IG light chain
61 also includes a variable domain that, once post-translationally associated with the heavy chain
62 variable domain in order to interact with cognate antigen; (3). Heavy chain constant region genes
63 separate IG into discrete isotypes (i.e., IgM, IgD, IgG, IgA, IgE). Each isotype demonstrates
64 unique effector properties that together represent the wide diversity of IG-mediated functions,
65 including binding of Fc receptors (FCR), activation of complement, opsonization, antibody-
66 dependent cellular cytotoxicity (ADCC) and antibody-dependent cellular phagocytosis (ADCP) (4,
67 5).

68 To recognize the wide range of pathogens humans encounter, the IG loci are highly
69 polymorphic and harbor diverse and complex sets of genes that recombine in each B cell to
70 encode up to 10^{13} unique IG specificities (6). B cells create this expansive catalog of specificities
71 through somatic recombination of the variable (V), diversity (D), and joining (J) genes in the IG
72 heavy chain locus (IGH), and V and J genes from the corresponding light chain loci, IG lambda
73 (IGL) and kappa (IGK) (7). First, a single D and J gene are recombined, while the intervening and
74 unused D and J gene sequences are removed by RAG recombinase (8). After D and J genes are
75 recombined, further recombination of a specific V gene to the DJ gene cassette completes the
76 formation of the full VDJ exon. After transcription of the recombined variable region, unique
77 constant region genes are spliced with the VDJ cassette to generate the heavy chain transcript
78 (7). Recombination of the light chain occurs similarly, recombining V and J genes only. Heavy

79 and light chain transcripts are independently translated and linked via covalent cysteine bonds
80 resulting in a fully functional IG protein prior to B cell cell-surface expression or secretion (**Figure**
81 **1A**) (9). Naïve B cells, which develop in the bone marrow from hematopoietic stem cell
82 progenitors, have undergone VDJ recombination but have not yet encountered antigen, and solely
83 express IgM and IgD (10). These naïve B cells then migrate to B cell zones in secondary lymphoid
84 tissues where they encounter antigen, driving further maturation and class switch recombination
85 (CSR) to enable the most effective humoral response (11). CSR allows for further excision of
86 constant genes at the DNA level, shifting the utilization of constant genes in the heavy chain locus
87 and resulting in different IG isotypes and subisotypes linked to the same VDJ (11).

88 The IgG isotype family is a composite of four subisotypes: IgG1, IgG2, IgG3, and IgG4.
89 Each IgG subisotype circulates at varied frequencies and facilitates unique immune functions. For
90 example, IgG1 is typically the most abundant circulating IgG and mediates proinflammatory
91 responses; IgG2 targets bacterial polysaccharides, conferring general protection from bacterial
92 pathogens; IgG3 confers protection against intracellular bacterial infections and enables clearing
93 of parasites; and IgG4 contains exclusive structural and functional characteristics often resulting
94 in anti-inflammatory and tolerance-inducing effects (5). Multiple studies have identified antibody-
95 mediated subisotype-specific pathogenicity in the context of autoimmune diseases and cancer
96 highlighting the need for further investigation of subisotype-specific IG repertoires (12-15).

97 Current AIRR-seq methods aim to resolve IG variable region transcripts to differing
98 extents. Profiling of this region, even in part, defines V, D and J gene usage while also providing
99 characterization of complementarity determining regions (CDR) 1, 2, and 3, which are
100 hypervariable and directly interact with target antigen (**Figure 1B**) (16). CDR3-targeted profiling
101 approaches, such as those used by Adaptive Biotechnologies (17), allow for V, D, and J gene
102 assignments but do not provide complete resolution of the entire variable region (17). Multiplexed
103 primer-based AIRR-seq strategies generate full variable region content but can be restrictive due
104 to the use of complex primer pools that can lead to inconsistent sequence coverage and will not

105 capture unknown genes and alleles for which there are no primers included. 5' RACE AIRR-seq
106 methods capture the full-length VDJ exon without multiplexed primer pools therefore limiting the
107 impact of primer bias and enabling discovery of novel genes and alleles (18). Crucially, 5' RACE
108 methods can also prime from the first constant region exon (CH1), allowing for determination of
109 isotype with the variable region. Additional methods have been developed that shift amplification
110 strategies by capturing more of the CH1 exon to enable subisotype resolution; however, these
111 methods sacrifice full and contiguous characterization of the V gene segment (19). All commonly
112 used AIRR-seq methods are further technically limited by the length restrictions (≤ 600nt) of short-
113 read next generation sequencing. As a result, no current AIRR-seq strategy resolves the complete
114 IG heavy chain transcript from the start of the constant exons to the end of the VDJ exon. Our
115 team has recently shown that population-based polymorphisms within the IGHV, IGHD, IGHJ and
116 IGL loci are far more extensive than previously known; the constant region has also been shown
117 to contain genomic diversity (20-22). Although it is understood that the Fc domain mediates IG
118 effector function, there is limited knowledge as to how genetic variation in this region may impact
119 functional capabilities (5, 23, 24). As such, there is a growing need to understand genomic
120 variation across the complete IG molecule. To address these limitations, we have developed
121 novel, end-to-end pipeline to target, profile, and characterize the IG heavy chain repertoire in the
122 context of isotype (IgG, IgM) and subisotype (IgG1, IgG2, IgG3, IgG4) groups.

123 Here, we present FLAIRR-seq, a novel targeted 5' RACE-based amplification of near full-
124 length IgG and IgM IGH transcripts, paired with single molecule real time (SMRT) sequencing,
125 resulting in highly accurate (mean read accuracy as high as Q60, 99.9999%), near full-length IG
126 sequences from RNA derived from whole blood, isolated PBMC, and purified B cells. When
127 analyzed with the Immcantation AIRR-seq tool suite, we demonstrate that FLAIRR-seq performs
128 comparably to standard 5'RACE AIRR-seq methods and single-molecule isoform sequencing
129 (IsoSeq) strategies for characterizing adaptive immune receptor repertoires. We further highlight
130 the novel features of FLAIRR-seq data, including phased identification of V, D, J, and constant

genes, facilitating the profiling of subisotype- and IGHC allele-specific repertoires and the characterization of class switch recombination.

157 **Materials and Methods**

158 **Sample collections**

159 Experiments were conducted using healthy donor peripheral blood mononuclear cells
160 (PBMC), purified B cells from healthy donors, or whole blood collected from hospitalized COVID-
161 19 patients (Supplementary Table 1). Commercially available healthy donor PBMC (STEMCELL
162 Technologies) and a subset of matched purified B cells were utilized to generate AIRR-seq and
163 FLAIRR-seq validation datasets. Full-length isoform sequencing (IsoSeq) was performed using B
164 cells isolated from the PBMC of a healthy, consented 57-year-old male donor at the University of
165 Louisville (UofL) School of Medicine. The UofL Institutional Review Board approved sample
166 collection (IRB 14.0661). For COVID-19 affected patient samples (n=5), whole blood was
167 collected from the Mount Sinai COVID-19 biobank cohort of hospitalized COVID-19 patients,
168 approved by the Institutional Review Board at the Icahn School of Medicine at Mount Sinai as
169 previously described (25).

170 **PBMC isolation and B cell purification**

171 Frozen healthy donor PBMCs were purchased, thawed, and aliquoted for use in
172 downstream experiments. For IsoSeq analyses, 175mL of venous blood was collected in a final
173 concentration of 6mM K_3EDTA using standard phlebotomy. PBMCs were isolated using Sepmate
174 PBMC Isolation Tubes (STEMCELL Technologies) as previously described (26), with an
175 additional granulocyte depletion step using the RosetteSep Human Granulocyte Depletion
176 Cocktail (STEMCELL Technologies) as directed by the manufacturer. B cells from the freshly
177 collected and frozen healthy donor PBMC were isolated using the EasySep Human Pan-B Cell
178 Enrichment Kit, as described by the manufacturer (STEMCELL Technologies). Briefly, B cells,
179 including plasma cells, were isolated by negative selection using coated magnetic particles. First,
180 the B cell enrichment cocktail was added to the sample and mixed for a 5-minute incubation at
181 room temperature, followed by addition of magnetic particles and further incubation for 5 minutes

182 on the benchtop. The sample tube was then placed on an EasySep magnet (STEMCELL
183 Technologies), and purified B cells were carefully eluted from the magnetic particles and
184 immediately used for RNA extraction.

185 **Genomic DNA and RNA extraction**

186 For the healthy frozen PBMC and matched purified B cells, genomic DNA (gDNA) and
187 RNA were co-extracted using the AllPrep DNA/RNA Mini Kit (Qiagen) according to the
188 manufacturer's instructions. For the freshly processed UofL healthy donor PBMC, purified Pan-B
189 cells were lysed in Buffer RLT Plus, and RNA was extracted with the RNeasy Plus Mini Kit
190 (Qiagen) per the manufacturer's protocol; no gDNA was collected from this sample. COVID-19
191 whole blood-derived RNA was extracted from samples collected in Tempus Blood RNA tubes
192 using the Tempus Spin RNA Isolation Kit (ThermoFisher) as described by the manufacturer. For
193 all samples, concentrations of RNA and gDNA (when appropriate) were assessed using the Qubit
194 4.0 fluorometer, with the RNA HS Assay Kit and Qubit DNA HS Assay Kit, respectively
195 (ThermoFisher Scientific). RNA and gDNA integrity were evaluated using the Bioanalyzer RNA
196 Nano Kit and DNA 1200 Kit, respectively (Agilent Technologies). Extracted RNA and gDNA were
197 stored at -80℃ and -20℃, respectively, until used.

198 **FLAIRR-seq targeted amplification of heavy chain transcripts**

199 Extracted RNA was thawed on ice and converted to first strand complementary DNA
200 (cDNA) using the SMARTer RACE 5'/3' Kit (Takara Bio USA), as described by the manufacturer
201 and a custom oligonucleotide that contained the template switch oligo and a unique molecular
202 identifier (5' TSO-UMI) for template switch during first strand cDNA synthesis. The following
203 reaction conditions were used: (i) a primary master mix was prepared with 4.0 µL 5X First-Strand
204 Buffer, 0.5 µL DTT (100 mM), and 1.0 µL dNTPs (20 mM) per reaction and set aside until needed;
205 (ii) in a separate 0.2-mL PCR tube, 10 µL of sample RNA and 1 µL 5'-CDS Primer A were
206 combined and incubated in a thermal cycler at 72°C (lid temperature: 105°C) for 3 minutes,

207 followed by cooling to 42°C for 2 minutes; (iii) after cooling, tubes were spun briefly to collect
208 contents and 1µL (12µM) of the 5' TSO-UMI was added to the RNA; (iv) 0.5 µL of RNase inhibitor
209 and 2.0 µL of SMARTScribe Reverse Transcriptase were added to the primary master mix tube
210 per sample and 8 µL of the combined master mix was then added to each RNA-containing sample
211 tube. First-strand cDNA synthesis reactions were incubated in a thermal cycler at 42°C (lid
212 temperature: 105°C) for 90 mins, followed by heat inactivation at 70°C for 10 minutes. Total first
213 strand cDNA generated in this reaction was diluted 1:2 with Tricine-EDTA Buffer before moving
214 onto targeted heavy chain transcript amplification.

215 To specifically amplify heavy chain transcripts from total first-strand cDNA, targeted IgG
216 and IgM transcript amplification reactions were performed using barcoded IgG (3' primer binding
217 in the constant region exon 3, CH3) or IgM (3' primer binding in the constant region exon 4, CH4)-
218 specific primers (Supplementary Table 2) and the following conditions: (i) 5 µL of diluted first-
219 strand cDNA was added to 0.2-mL PCR tubes; (ii) a master mix was generated using 10 µL 5X
220 PrimeSTAR GXL Buffer, 4 µL GXL dNTP mixture, 28 µL PCR-grade water, 1 µL PrimeSTAR GXL
221 Polymerase and 1 µL 10x UPM form the SMARTer RACE 5'/3' Kit per reaction; (iii) 44 µL master
222 mix was added to each reaction tube followed by 1 µL of the appropriate barcoded IgG (CH3) or
223 IgM (CH4) primer. Different temperatures were used for annealing of IgG- (63.5°C) and IgM-
224 specific primers (60°C) to account for primer specific melting temperatures and to enhance
225 targeted amplification specificity. Amplification conditions for full-length IgG were: 1 minute at
226 95°C, followed by 35 amplification cycles of 95°C for 30 sec, 63.5°C for 20 sec, and 2 minutes at
227 68°C, followed by a final extension for 3 minutes at 68°C and hold at 4°C. Amplification conditions
228 for full-length IgM were: 1 minute at 95°C, followed by 35 amplification cycles of 95°C for 30 sec,
229 60°C for 20 sec., and 2 minutes at 68°C, followed by a final extension for 3 minutes at 68°C and
230 hold at 4°C. Final amplification reactions were purified using a 1.1x (vol:vol) cleanup with ProNex
231 magnetic beads (Promega) . Successfully amplified products were quantified with Qubit dsDNA

232 HS assay (ThermoFisher Scientific) and length was evaluated with the Fragment Analyzer
233 Genomic DNA HS assay (Agilent). Samples were equimolar pooled in 8-plexes for SMRTbell
234 library preparation and sequencing.

235 **FLAIRR-seq SMRTbell library preparation and sequencing**

236 Eight-plex pools of targeted IgG or IgM amplicons were prepared into SMRTbell
237 sequencing templates according to the "Procedure and Checklist for Iso-Seq Express Template
238 for Sequel and Sequel II systems" protocol starting at the "DNA Damage Repair" step and using
239 the SMRTbell Express Template Prep Kit 2.0, with some modifications (Pacific Biosciences).
240 Briefly, targeted IgG and IgM amplicons underwent enzymatic DNA damage and end repair,
241 followed by ligation with overhang SMRTbell adapters as specified in the protocol. To increase
242 consistency in SMRTbell loading on the Sequel IIe system, we further treated the SMRTbell
243 libraries with a nuclease cocktail to remove unligated amplified products, using the SMRTbell
244 Enzyme Cleanup Kit, as recommended by the manufacturer (Pacific Biosciences). Briefly, after
245 heat-killing the ligase with an incubation at 65°C, samples were treated with a nuclease cocktail
246 at 37°C for 1 hour, and then purified with a 1.1X Pronex cleanup. Final SMRTbell libraries were
247 evaluated for quantity and quality using the Qubit dsDNA HS assay and Fragment Analyzer
248 Genomic DNA assay, respectively. Sequencing of each 8-plex, barcoded sample pool was
249 performed on one SMRTcell 8M using primer v4 and polymerase v2.1 on the Sequel IIe system
250 with 30 hr movies. Demultiplexed, high-fidelity circular consensus sequence reads ("HiFi reads")
251 were generated on the instrument for downstream analyses.

252 **AIRR-seq SMARTer Human BCR IgG/IgM IGH sequencing**

253 Matched healthy donor RNA was used to generate targeted IgG and IgM AIRR-seq
254 libraries using the SMARTer Human BCR IgG IgM H/K/L Profiling Kit (Takara Bio USA) according
255 to the manufacturer's instructions with no modifications. Briefly, for each sample, proprietary IgG
256 and IgM primers were used to amplify heavy chain transcripts following a 5'RACE reaction. AIRR-

257 seq libraries were then quality controlled using the 2100 Bioanalyzer High Sensitivity DNA Assay
258 Kit (Agilent) and the Qubit 3.0 Fluorometer dsDNA High Sensitivity Assay Kit. Sequencing on the
259 MiSeq platform using 300bp paired-end reads was performed using the 600-cycle MiSeq Reagent
260 Kit v3 (Illumina) according to the manufacturer's instructions, and FASTQ reads were generated
261 using the associated DRAGEN software package (Illumina).

262 **B cell IsoSeq**

263 RNA extracted from healthy sorted B cells was used to generate IsoSeq SMRTbell
264 libraries following the "Procedure & Checklist Iso-Seq Express Template Preparation for the
265 Sequel II System" with minor adaptations compared to the manufacturer's instructions. Briefly,
266 Iso-Seq libraries were generated using 500 ng high-quality (RIN > 8) RNA as input into oligo-dT
267 primed cDNA synthesis (NEB). Barcoded primers were incorporated into the cDNA during second
268 strand synthesis. Following double-stranded cDNA amplification, transcripts from two samples
269 sourced from purified B cells and NK cells were equimolar pooled as previously described (26).
270 SMRTbells were generated from the pooled cDNA as described above for the FLAIRR-seq
271 amplification products, including the addition of a nuclease digestion step. Quantity and quality of
272 the final IsoSeq libraries were performed with the Qubit dsDNA High Sensitivity Assay Kit and
273 Agilent Fragment Analyzer Genomic DNA assay, respectively. This 2-plex IsoSeq pool was
274 sequenced using primer v4 and polymerase v2.1 on the Sequel IIe system with a 30-hour movie.
275 HiFi reads were generated on instrument before analyses. Demultiplexing of barcoded samples
276 and generation of full-length non-concatemer (FLNC) predicted transcripts were performed using
277 the IsoSeq v3 pipeline available through SMRTLink (v.10.2). B-cell-derived FLNC reads were
278 mapped to the human genome using the GMAP reference database, and reads derived from
279 chromosome 14 were extracted for downstream IGH transcript characterization via Immcantation,
280 as described below.

281 **Immcantation analyses of IgG and/or IgM IGH variable region repertoires**

282 Analyses of FLAIRR-seq, AIRR-seq, and Iso-seq datasets were performed using
283 Immcantation tools to characterize IG variable regions (27, 28). Demultiplexed barcoded HiFi (for
284 SMRT sequencing data) or FASTQ (for AIRR-seq) reads were first processed using the pRESTO
285 tool for quality control, UMI processing, and error profiling (28). For AIRR-seq, pRESTO analysis
286 data from paired-end reads ("R1" and "R2") were trimmed to remove bases with < Q20 read
287 quality and/or <125 bp length using the "FilterSeq trimqual" and "FilterSeq length", respectively.
288 IgG and IgM CH3 or CH4 primer sequences were identified with an error rate of 0.2, and primers
289 identified were then noted in FASTQ headers using "MaskPrimers align". Next, 12 basepair (bp)
290 UMIs were located and extracted using "Maskprimers extract". Sequences found to have the
291 same UMIs were grouped and aligned using "AlignSets muscle," with a consensus sequence
292 generated for each UMI using "BuildConsensus". Mate pairing of AIRR-seq reads was conducted
293 using a reference-guided alignment requiring a minimum of a 5 bp overlap via "AssemblePairs
294 sequential". After collapsing consensus reads with the same UMI ("conscount") using
295 "CollapseSeq"," reads with < 2 supporting sequences were removed from downstream analysis.
296 For pRESTO processing of FLAIRR-seq, single HiFi reads ("R1") reads did not require trimming
297 due to > Q20 sequence quality across all bases. 5'TSO-UMI and CH3 or CH4 region primers were
298 identified along with a 22bp UMI with an error rate of 0.3 using "MaskPrimers align". Reads were
299 then grouped and aligned using "AlignSets muscle". Due to the single molecule nature of FLAIRR-
300 seq reads, no mate pairing was required. Consensus reads were then generated as described
301 above, including removal of sequences with < 2 supporting reads.

302 pRESTO-filtered reads for both AIRR-seq and FLAIRR-seq data were then input into the
303 Change-O tool (Table 1). Iso-seq required no initial processing from pRESTO and was input into
304 Change-O for IG gene reference alignment along with AIRR-seq and FLAIRR-seq data using
305 "igblastn", clonal clustering using "DefineClones", and germline reconstruction and conversion
306 using "CreateGermlines.py" and the GRCh38 chromosome 14 germline reference (27). Fully

307 processed and annotated data was then converted into a TSV format for use in downstream
308 analyses. The Alakazam Immcantation tool suite was then used to quantify gene usage analysis,
309 calculate CDR3 length, assess somatic hypermutation frequencies, and analyze clonal diversity
310 (27). Scoper clonal assignment by spectral clustering was conducted for COVID-19 patient time
311 course samples (29, 30). For clonal lineage tree analysis, the Dowser tool was used to examine
312 clonal diversity and CSR over time (31).

313 **Targeted IG gDNA capture, long-read sequencing, and IGenotyper analyses.**

314 FLAIRR-seq validation samples also underwent IGHC targeted enrichment and long-read
315 sequencing as previously described (32). Briefly, gDNA was mechanically sheared and size
316 selected to include 5-9kb fragments using the BluePippin (Sage Science). Samples were then
317 end repaired and A-tailed using the standard KAPA library preparation protocol (Roche).
318 Universal priming sequences and barcodes were then ligated onto samples for multiplexing
319 (Pacific Biosciences). Barcoded gDNA libraries were captured using IGH-specific probes
320 following a SeqCap protocol (Roche). 26 IGH-enriched samples were purified and pooled together
321 for SMRTbell library prep as described above, including the final nuclease digestion step. Pooled
322 SMRTbells were annealed to primer v4, bound to polymerase v2.0, and sequenced on the Sequel
323 IIe system with 30h movies. After sequencing, HiFi reads were generated and analyzed by the
324 IGenotyper pipeline (32). In brief, IGenotyper was used to detect single nucleotide variants and
325 assemble sequences into haplotype assemblies for downstream constant region genotyping.

326 **Constant region genotyping with FLAIRR-seq**

327 To genotype constant region genes and alleles from FLAIRR-seq data, productive reads
328 were filtered by constant region length (900bp–1100bp) and aligned to the chromosome 14 hg38
329 reference using minimap2 along with SAMtools to generate sorted and indexed bam files (33,
330 34). WhatsHap was used to identify, genotype, and phase single nucleotide variants (SNV) (35).
331 Phased SNVs were used to assign each read to a haplotype using MsPAC. Reads from each

332 haplotype and gene were clustered using CD-HIT using a 100% identity clustering threshold
333 parameter, and a single representative read from the largest cluster was BLASTed to the
334 ImMunoGeneTics (IMGT) reference database (downloaded on 2/21/22) to determine the constant
335 region allele usage (36, 37). The representative read was selected based on 100% identity to all
336 other sequences in that cluster. Constant region alleles observed by FLAIRR-seq were
337 designated as “novel” if there was no match to the IMGT database or “novel, extended” if a match
338 was detected to a partial allele found in IMGT, but the IMGT allele was a substring of the FLAIRR-
339 seq identified allele. In cases for which FLAIRR-seq resulted in ambiguous allele calls,
340 visualization of IGenotyper DNA haplotype assemblies and FLAIRR-seq data were performed
341 using the Integrated Genomics Viewer (IGV v2.12.3) software to manually confirm allele
342 assignment based on IGenotyper ground truth.

343

344

345

346

347

348

349

350

351

352

353

354

355 **Results**

356 **Validation and benchmarking of the FLAIRR-seq method**

357 Current methods for AIRR-Seq utilize targeted amplification of the variable region and, in
358 some cases, a small portion of the first constant region exon (CH1), in conjunction with short-read
359 sequencing to characterize IG repertoires. However, this minimal examination of CH1 is primarily
360 used to define IG isotype groups. No current method defines both the heavy chain variable and
361 constant regions with allele-level resolution allowing for subisotype classification. To address
362 these technical limitations, we developed the FLAIRR-Seq method (Figure 1C), a targeted 5'
363 RACE approach combined with SMRT sequencing to generate highly accurate, near full-length
364 IgG (~1500 bp) and/or IgM (2000 bp) sequences, allowing for direct, simultaneous analysis of the
365 heavy chain variable and constant regions, including gene/allele identification for V, D, J and C
366 segments, and isotype- and subisotype-specific repertoire profiling.

367 To evaluate and validate the capabilities of FLAIRR-seq, matched FLAIRR-seq and AIRR-
368 seq analyses were performed on ten healthy donor PBMC samples. FLAIRR-seq data were
369 filtered from the initial HiFi reads (>Q20) to include only >Q40 reads. The average read quality of
370 these filtered reads was >Q50 (99.999%), with a pass filter rate ranging from 88%-93% of total
371 reads. AIRR-seq FASTQ reads were filtered for those with an average quality of >Q30 (99.9%).
372 These filtered reads were used as input into the Immcantation suite, specifically the pREST-O
373 and Change-O tools, for V, D, and J gene assignment, and different repertoire feature analyses,
374 such as identification of clones, extent of somatic hypermutation, and evaluation of CDR3 lengths.
375 As shown in **Table 1,** fewer overall FLAIRR-seq reads were used as input into the Immcantation
376 analyses, after required filtration and read assembly steps (which were not needed for the high-
377 quality single-molecule FLAIRR-seq reads). FLAIRR-seq resulted in comparable or, in many
378 cases, increased number of unique V, D, and J gene allele pairs (VDJ), clones, and CDR3
379 sequences identified compared to the matched AIRR-seq-derived samples in both the IgG and

380 IgM repertoires (schematized in Figure 2A). These basic sequencing and initial analysis metrics
381 demonstrated that FLAIRR-seq could produce high-quality variable region data for detailed IG
382 repertoire analyses. Furthermore, the ability to utilize Immcantation shows that current AIRR-seq
383 tools are capable of analyzing FLAIRR-seq data. While input RNA mass used for FLAIRR-seq
384 was often more than for AIRR-seq, ongoing and future optimization of the method is aimed at
385 reducing RNA input. Furthermore, a comparative cost analysis was performed. To obtain the
386 closest metric to a direct comparison, we calculated the cost per "actionable read", defined as the
387 read number per sample and per method after read and length quality filtration and assembly, but
388 prior to cluster consensus, performed in the pRESTO pipeline. This method was used to represent
389 the total unique single molecule or assembled variable region templates captured by either the
390 FLAIRR-seq or AIRR-seq methods that passed all necessary quality control criteria for
391 downstream annotation and analyses irrespective of biologic repertoire diversity or clonality,
392 removing technology/platform biases. Lastly, to remove the impact of pooling differences, we
393 used this "per actionable read" cost to calculate the cost for the generation of 15,000 "actionable
394 reads" as our standard price. For AIRR-seq this cost was $25.50 per sample, whereas for
395 FLAIRR-seq, this cost was $33.57 per sample. These costs reflect needed reagents and
396 consumables only, assume instrumentation access and do not include labor. Future optimization
397 for FLAIRR-seq will include integrating a multiplexed array sequencing (MAS) step to concatenate
398 reads and enhance overall depth of sequencing and multiplexing capacity per pool, resulting in
399 decreasing costs per sample (38).

400 While optimizing FLAIRR-seq sample preparation, we examined whether upstream
401 isolation of B cells before FLAIRR-seq molecular preparation would enhance the ability to detect
402 unique variable region gene usage. To do this, aliquots of PBMC (n=4) were split into two groups
403 for RNA extraction: (i) RNA derived from bulk PBMC, and (ii) RNA isolated from purified pan B-
404 cells, followed by FLAIRR-seq preparation, SMRT sequencing, and Immcantation analysis of both

405 groups. V gene usage correlations between groups are shown in Figure 2B and demonstrate a
406 significant association (p-values ranging from 0.033 to $4.1e^{-16}$) strongly supporting the conclusion
407 that B cell isolation before RNA extraction was not necessary to achieve comparable gene usage
408 metrics. The limited differences that were observed could be explained by template sampling
409 differences between the two experiments. Due to the strong associations observed and the ease
410 of processing PBMC in bulk, we moved forward with bulk-derived PBMC material for the
411 remainder of our analyses.

412 We established FLAIRR-seq performance by comparing its output to the commonly used
413 5' RACE AIRR-seq method. 5' RACE AIRR-seq was chosen as it provides the most extensive
414 AIRR resolution of standard methods in the field, capturing both the variable region and a small
415 portion of the constant region CH1 exon, allowing for isotype differentiation. Following matched
416 preparation of both FLAIRR-seq and AIRR-seq libraries from healthy donor PBMC samples
417 (n=10), we compared many variable region features to benchmark FLAIRR-seq performance.
418 First, we evaluated variable region gene usage frequencies. We observed strongly significant
419 associations (p-values ranging from 0.033 to $<2.2e^{-16}$) between V, D and J gene usage for both
420 IgG and IgM isotypes (Figure 2C), indicating that FLAIRR-seq accurately and comparably
421 resolves variable region genes. To note, J genes showed less significance (larger p-value) across
422 all comparisons due to the relatively few genes that make up the J gene family compared to the
423 more diverse V and D families. We next investigated the performance of both methods in terms
424 of resolving CDR3 lengths (Figure 2C) and somatic hypermutation frequencies contained in
425 productive sequences (Figure 2E), which are often used as measures of evaluating B cell affinity
426 maturation. FLAIRR-seq and AIRR-seq performed robustly, with significantly longer CDR3
427 regions detected with FLAIRR-seq across all samples in both the IgM and IgG groups. The
428 characterization of unusually long CDR3 regions (> 40 nt) in the IgG sequences with FLAIRR-seq
429 that were not observed in the matched AIRR-seq data, could be due the higher contiguity and

430 quality afforded from longer read length when using FLAIRR-seq. Single-molecule resolution of
431 near full-length heavy chain transcripts would allow for the inclusion of unusually long variable
432 regions that may have been missed by short-read 2x 300 bp paired-end sequencing strategies if
433 reads 1 and 2 did not maintain enough overlap in this region to assemble. Overall, these data
434 demonstrate that FLAIRR-seq achieves unambiguous and comparable, if not improved,
435 resolution of the variable region genes and features, and improved in terms of CDR3 length
436 resolution, and can be used with publicly available AIRR-seq analyses tools to provide commonly
437 used metrics for repertoire analysis.

438 Others have recognized the power of long-read sequencing to resolve B cell repertoires
439 but have used bulk Iso-Seq methods to examine full-length transcripts from isolated B cells (39).
440 The IsoSeq method captures full-length transcripts expressing a poly(A) tail without bias through
441 oligo dT-based priming. The trade-offs of this approach are throughput, depth, and cost, as
442 IsoSeq processing generates a complete transcriptome per sample without enrichment of heavy
443 chain sequences, which then need to be filtered out and analyzed, resulting in a considerable
444 amount of non-repertoire data to be discarded. To investigate whether the untargeted
445 transcriptome-wide IsoSeq method would resolve a qualitatively different repertoire than FLAIRR-
446 seq, which would have indicated FLAIRR-seq-driven primer bias, we performed matched Iso-Seq
447 and FLAIRR-seq on purified B-cell derived RNA. Variable gene usage frequencies between
448 IsoSeq- and FLAIRR-seq-derived V, D, and J calls were examined by Spearman ranked
449 correlation and showed highly significant associations between methods (V genes: r^2 = 0.94,
450 $p=2.2e^{-16}$; D genes: r^2 = 0.92, p =$1.4e^{-13}$; J genes: r^2 = 1.0, p = 0.0028). These data strongly
451 suggest that FLAIRR-seq has very limited to no primer-driven bias compared to whole
452 transcriptome data. Collectively, this benchmarking dataset confirmed that FLAIRR-seq is robust
453 and accurate, providing complete characterization of IG variable genes and associated commonly
454 used IG repertoire metrics, with limited increases in per sample cost.

455 **IGenotyper and FLAIRR-seq provide constant region gene and allele identification**

456 Although confirmation that FLAIRR-seq successfully characterized IG variable region
457 transcripts, the true value added by FLAIRR-seq is the novel capabilities it provides in resolving
458 the constant region, including gene and allele usage, subisotype identification, and phasing of
459 variable and constant regions for comprehensive repertoire analysis**.** To evaluate the capabilities
460 and accuracy of constant region gene and allele identification with FLAIRR-seq, we first
461 established a ground truth dataset of IGHC alleles for all 10 samples by targeted sequencing of
462 the germline IGH locus (Figure 3A) using IGenotyper, as previously described (32). Constant
463 region alleles called by IGenotyper were assigned to one of three categories, schematized in
464 Figure 3B: (i) “exact match” - known constant region alleles found in the IMGT database; (ii)
465 “novel not in IMGT” – a constant region allele not found in the IMGT database; or (iii) “extended”
466 – constant region alleles that appeared to match partial alleles in the IMGT database, but
467 IGenotyper data extended the sequence beyond what has been previously characterized.
468 IGenotyper identified a total of 32 alleles across all individuals and classifications, as schematized
469 in Figure 3C. Furthermore, we discovered IGHG4 gene duplications in in 4/10 samples.
470 Identification of these constant region gene duplications further highlights the structural
471 complexity and diversity of IGH.

472 Given the extensive additional constant region gene content resolved by IGenotyper
473 compared to IMGT, we used the IGenotyper-derived IGHC gene database as the ground-truth for
474 evaluation of the capabilities of FLAIRR-seq data to identify and resolve constant region alleles.
475 FLAIRR-seq data alone identified a total of 23/32 (72%) alleles at 100% identity with no additional
476 alleles (false-positives) detected as compared to the IGenotyper ground-truth database (Figures
477 3D and E). Due to the internal placement of the CH3- and CH4-specific primers used to target IG
478 transcripts, the entirety of these exons were not resolved and, in some cases, resulted in slightly
479 truncated transcripts that resulted in ambiguous allele calls when aligned to our ground-truth

480 IGenotyper constant region database (Figure 3C). For example, in IGHG1 alleles the complete
481 sequence of the CH3 exon was often necessary to distinguish between two closely related alleles,
482 as well as determine if an individual was homozygous or heterozygous at these positions. In these
483 cases, FLAIRR-seq data were able to substantially narrow the field of possible alleles at that
484 position in an individual, but the IGenotyper data was used to identify the exact allele. Using
485 FLAIRR-seq data alone also limited phasing capabilities for some constant region genes due to
486 missing allele-defining SNVs in the truncated transcripts resulting in false homozygous calls,
487 primarily in IGHG1 genes (30%). Further improvements to FLAIRR-seq will aim to iterate in primer
488 design to capture the remaining 3' region of CH3 to increase genotype capabilities of FLAIRR-
489 seq as a stand-alone method.

490 In total, IGenotyper data identified 32 constant region alleles across the 10 healthy donors
491 examined, including 11 alleles (34%) that were completely novel and 17 (53%) alleles that were
492 extended compared to those that were curated in IMGT at the time of publication. FLAIRR-seq
493 was also able to resolve 23/32 (72%) these alleles when assessed as a stand-alone genotyping
494 method. Given the relatively small size of this proof-of-concept healthy donor cohort, the
495 identification of 28 (87%) novel or novel-extended alleles here underscores the extensive
496 polymorphism in this region, despite a historical assumption that these genes are "constant" and
497 reflects the paucity of information regarding this locus in existing immunogenomics databases.

498 **FLAIRR-seq enables subisotype-specific repertoire analyses**

499 Once subisotypes were identified, we performed subisotype-specific repertoire analyses
500 (Figure 4). We initially examined V gene family usage as a function of IgG subisotype across all
501 10 healthy donors. From this analysis we observed V gene family usage diversity within each
502 subisotype across all seven V gene families (Figure 4A). For example, significantly disparate V
503 family gene usage patterns are seen across subisotypes for the IGHV1 genes where IGHV1
504 genes are most highly expressed in the IgG4-specific repertoire and are seen the least within the

505 IgG2-specific repertoire (ANOVA $p=1.3e^{-07}$). We chose to investigate subisotype-specific V, D,
506 and J gene usage for samples 1008 and 602c in detail as a demonstration of this capability.
507 Characterization of constant region genes and alleles within FLAIRR-seq data allowed for the
508 relative frequency of heavy chain transcripts per isotype class, subisotype, and subisotype-
509 specific allele to be assigned. Once classified by subisotype, detailed analysis of variable region
510 repertoire per subisotype-specific allele was possible (Figure 4B). For samples 1008 and 602c,
511 we observed that IGHV3 genes were highly expressed across all subisotypes, whereas IGHV1
512 genes are relatively elevated in the IgG4 population compared to the IgG1, IgG2, and IgG3
513 compartments (Figure 4C and D). However, samples 1008 and 602c differ in their expression of
514 IGHV1 family genes, which are expressed at relatively higher levels in sample 602c compared to
515 sample 1008 (Figure 4C and D). In the context of standard AIRR-seq analyses, we would not be
516 able to identify allele-resolved V gene usage or enrichment within subisotype populations, which
517 could be essential for linking subisotype functionality with VDJ clonotypes.

518 **FLAIRR-seq identifies subisotype-specific clonal expansion and CSR in longitudinal**
519 **samples.**

520 It was critical to demonstrate the feasibility of performing FLAIRR-seq from clinically
521 relevant sample types, such as whole blood. To do this, we evaluated FLAIRR-seq performance
522 in a small subset of samples collected in April 2020 from hospitalized COVID-19 patients (n=5)
523 from the Mount Sinai COVID-19 biobank cohort at the Icahn School of Medicine at Mount Sinai
524 (25). RNA was directly extracted from whole blood, as outlined in the Materials and Methods, and
525 robust FLAIRR-seq amplification of targeted IgG and IgM repertoires was observed from all
526 samples (data not shown).

527 IG responses are highly dynamic, with specific IG clones expanding upon activation by
528 cognate antigen. We were interested to know if these dynamics could be captured by FLAIRR-
529 seq, given the capability to identify clonotypes with the subisotype resolved. We had the

530 opportunity to evaluate FLAIRR-seq resolved repertoires over time in four samples collected from
531 one individual over their >13-day hospitalization. Patient PICR7356 was hospitalized with severe
532 COVID-19 disease, and blood draws were taken on days 1, 4, 8, and 13 post-hospitalization
533 (Figure 5A) and analyzed with FLAIRR-seq across all time points. After initial FLAIRR-seq
534 processing and analysis, we defined unique clones using SCOPer, which clusters groups based
535 on CDR3 similarity and mutations in V and J genes and provided clone counts across the four
536 timepoints examined. Overall, we observed that most clones were expressed within the IgG1
537 compartment, but longitudinal clone distribution across subisotypes fluctuated over time.
538 Specifically, the IgG3-specific repertoire expanded from day 1 to day 4, but then contracted in
539 overall frequency from day 8 to day 13 (Figure 5B). To assess clonal diversity within each
540 subisotype repertoire across time, we calculated the Simpson's diversity index (q=2) using
541 Alakazam (27). All subisotype-specific repertoires became less diverse from day 1 to day 4,
542 suggesting clonal expansion across the IgG repertoire (Figure 5C). To note, IgG4 was not
543 included in diversity calculations because IgG4 would have required higher sequencing depth to
544 ascertain diversity in this individual, given the overall lower expression of IgG4 transcripts in this
545 individual. Subisotype-specific IG repertoire polarity was also calculated by counting the number
546 of clones needed to represent 80% of the total repertoire and then expressing this number as a
547 percentage of the total (Figure 5D). In this metric, the lower the clone number that is needed to
548 represent 80% of the repertoire, the higher the extent of clonal expansion observed. Polarity
549 values confirmed those seen with the diversity index, demonstrating an increase of clonal
550 expansion (i.e., decreased polarity) at day 4 across IgG1, IgG2, and IgG3 compartments, which
551 returned to baseline at later timepoints. Notably, although expressed at low levels, polarity of the
552 IgG4-specific repertoire was examined and was an exception to this trend, demonstrating
553 indications of clonal expansion at day 8 and day 13.

554 CSR is a critical immunological phenomenon allowing for the replacement of constant
555 region genes with others and, therefore, altering associated effector functions while maintaining
556 antigen-specific variable regions. Given the ability of FLAIRR-seq to resolve variable clonotypes
557 and constant region gene usage, we evaluated whether this method could identify signatures of
558 CSR in a longitudinal repertoire. To do this, we examined clones shared across timepoints and
559 asked how these clones were distributed within each subisotype-specific repertoire over time. In
560 total, 34 unique clonal lineages were shared across all timepoints and could be examined for
561 signatures of CSR. As a proof-of-concept, we examined the relative frequency of the dominant
562 clone 13611 within each subisotype repertoire across all time points (Figure 5E). To establish that
563 we were examining possible CSR, we confirmed that all 13611 clonal lineage sequences shared
564 the same variable region alleles. On day 1 post-hospitalization, clone 13611 sequences were
565 found equally distributed within the IgM and IgG2 repertoire, indicating that these B cells had not
566 undergone complete maturation. By day 4, we see clear evidence of B cell maturation and CSR
567 by the dominance of IgG2 usage by clone 13611 sequences. Residual low-level IgM sequences
568 rebounded on day 8, while also showing low-level CSR from IgM to IgG1 at that same time point.
569 To visualize these dynamics, we built a phylogeny using the Dowser pipeline via Immcantation.
570 Highlighted in the red box on the phylogenetic tree shown in Figure 5F, we observe a clustered
571 IgG1 transcript emerging from IgM-specific clade, indicating CSR from IgM to IgG1. Together,
572 these data provide demonstrative proof-of-concept evidence that FLAIRR-seq profiling performs
573 robustly on clinical samples, including RNA directly extracted from whole blood. In addition, these
574 data provide novel repertoire resolution than what would have been provided with standard AIRR-
575 seq methods, including analysis of subisotype-specific repertories, evaluation of clonal
576 expansion, and characterization of CSR in the IgG and IgM repertoires.

577

578 **Discussion**

579 Here we present the development, validation, and application of FLAIRR-seq, a novel
580 method to resolve near full Ig IGH transcripts from bulk PBMC-, isolated B cell- and whole blood-
581 derived total RNA. FLAIRR-seq enables highly accurate, simultaneous resolution of Ig variable
582 and constant regions and suggests that constant region polymorphism is far more extensive than
583 previously assumed. FLAIRR-seq performed equivalent to or with increased resolution compared
584 to existing standard 5'RACE AIRR-seq methods when resolving variable gene calls, CDR3
585 lengths, and somatic hypermutation signatures, suggesting that our CH3/CH4 targeting strategies
586 did not compromise variable region characterization while simultaneously adding the capability to
587 resolve extensive constant region genomic variation. Little to no primer bias was observed when
588 compared to IG repertoire profiling from total mRNA Iso-seq methods. FLAIRR-seq provides the
589 novel ability to use constant region gene usage to identify IG subisotypes and genotype heavy
590 chain transcripts, linking these data back to evaluate subisotype-specific repertoires, clonal
591 expansion, and CSR. Underscoring the underappreciated extent of constant region genomic
592 variation, our FLAIRR-seq profiling of a restricted cohort of only 10 individuals from relatively
593 homogenous backgrounds still identified 4 and 6 completely novel constant region alleles in IgM
594 and IgG, respectively, and extended an additional 16 alleles beyond which had been available in
595 the IMGT database.

596 The unique capabilities of FLAIRR-seq will allow for novel examination of IG repertoires,
597 including the characterization of variable gene usage and clonotype distribution within unique
598 subisotype subsets. This perspective has the potential to provide key insights into dynamic IG
599 responses in diseases known to be mediated by subisotype-specific IG or have skewed
600 subisotype distribution as predictive markers of disease, including Myasthenia gravis (mediated
601 by pathogenic autoantibodies across subisotypes that give rise to varied disease pathologies),
602 Acute Rheumatic Fever (associated with elevated IgG3), and melanoma (skewing towards IgG4
603 in late-stage disease thought to be indicative of tolerogenic responses and poor prognosis)(12-

604 14). These subisotype-specific repertoire profiling approaches may be the first step toward
605 identification of unique IG clones that mediate disease pathogenicity or serve as high resolution
606 biomarkers to disease progression, as well as open the door for potential functional experiments
607 on subisotype clones of interest, including examining the functional impact of the novel constant
608 region alleles identified here. Expanded population-based FLAIRR-seq profiling and curation of
609 novel constant region alleles, particularly in conjunction with IGenotyper targeted genomic
610 assembly efforts in IGH, will be a significant first step in defining the full extent of variation in a
611 region too long assumed to be relatively invariant.

612 The Fc region is known to be critical for modulating differential IG effector functions. These
613 differential functionalities are currently understood to be regulated by differential posttranslational
614 modifications, such as variable glycosylation (40-43). Future FLAIRR-seq profiling will be a
615 valuable tool to investigate how genomic variation across constant region genes impacts residue
616 usage and resultant Fc receptor binding, signaling potential, crosslinking, and potential for
617 posttranslational modification, all of which would be expected to alter downstream effector
618 function potential, such as ADCC, ADCP, and complement fixation.

619 We further demonstrate that FLAIRR-seq can effectively examine clonal expansion and
620 CSR in longitudinal samples, demonstrating the feasibility of using FLAIRR-seq to resolve IG
621 repertoire dynamics. This increased resolution will further our understanding of IG repertoire
622 evolution in the transition of acute to chronic disease states, many of which are associated with
623 overall IgG subisotype distribution changes that are thought to reflect the inflammatory milieu (14,
624 44). One example is advanced melanoma, where late-stage disease is characterized by elevated
625 IgG4 compared to IgG1, which is believed to reflect a more tolerizing, pro-tumor environment (12).
626 FLAIRR-seq examination of these samples may identify specific IG distribution patterns that could
627 act as biomarkers of disease progression. Moving forward it is critical to account for all variability
628 within the IG repertoire for the most comprehensive understanding of repertoire dynamics and

629 the myriad factors impacting antibody effector function. Together, the data presented here
630 demonstrate that the FLAIRR-seq method provides a comprehensive characterization of allele-
631 resolved IgG and IgM repertoires, detailing variable region gene usage and measurements of
632 maturation, isotype and subisotype identification, and the unappreciated extent of constant region
633 variation, which will be necessary to fully appreciate the impact of IG genomic variation in health
634 and disease.

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652 **Figure Legends:**

653 **Figure 1. Overview of IG structure and FLAIRR-seq molecular method.** (A) Schematic
654 representation of the IGH locus, heavy chain transcript structure, and functional IG protein. (B)
655 Comparative coverage across the heavy chain transcript of commonly used AIRR-seq methods
656 compared with FLAIRR-seq. (C) FLAIRR-seq molecular pipeline: RNA (brown) was converted to
657 first-strand cDNA (red) using the 5' RACE method, incorporating a 5' TSO-UMI (pink) via template
658 switch. Second strand amplification specifically targeted IgG and IgM molecules through priming
659 of the 5' TSO-UMI and the 3' constant region IGH exon 3 (CH3) for IgG, or CH4 for IgM. A 16bp
660 barcode was incorporated into the 3' CH3/CH4 primers to enable sample multiplexing post-
661 amplification. (D) Fragment Analyzer traces of final FLAIRR-seq products at ~1500bp (IgG) and
662 ~2000bp (IgM).

663 **Figure 2. FLAIRR-seq shows robust characterization of IG variable regions.** (A) Schematic
664 representation analyses metrics evaluated including clones, unique VDJ, CDR3 length, and
665 somatic hypermutation. (B) Spearman ranked correlations and p-values of V, D, and J gene usage
666 frequencies identified by FLAIRR-seq performed on matched total PBMC and purified B cells. (C)
667 Heatmap of Spearman ranked correlations and p-values of V, D and J gene usage frequencies
668 between FLAIRR-seq and AIRR-seq processed samples. (D) Boxplots of CDR3 length defined
669 by FLAIRR-seq or AIRR-seq analysis in both IgM- and IgG-specific repertoires. (E) Boxplots of
670 somatic hypermutation frequencies defined by FLAIRR-seq or AIRR-seq in IgM- and IgG-specific
671 repertoires. (F) Spearman ranked correlation of V, D, and J gene usage frequencies between
672 FLAIRR-seq-based and Iso-Seq-based repertoire profiling. Significant differences between
673 FLAIRR-seq and AIRR-seq data indicated by * ($p < 0.05$) or ** ($p < 0.01$)

674 **Figure 3. FLAIRR-seq provides novel constant region resolution for allelic discovery.** (A)
675 Overview of experimental design and pipeline overviews of genotyping by IGenotyper (gDNA)
676 and FLAIRR-seq (RNA). (B) Schematic depicting constant region allele types identified by

677 IGenotyper, including identification of (i) exact match to IMGT, (ii) novel, not in IMGT and (iii)
678 extended. (C) Pie chart and stacked bar graph representing total alleles and their types (exact
679 match, novel, not in IMGT, or extended) per IGHC gene as identified by IGenotyper. * indicates
680 additional allele found due to IGHG4 duplicaton (D) Table summarizing of novel (pink) and
681 extended (diamond) alleles resolved in FLAIRR-seq data alone. E) Pie chart and stacked bar
682 graph of total alleles and their types (exact match, novel, not in IMGT, extended, or ambiguous)
683 resolved by FLAIRR-seq data alone.

684 **Figure 4. FLAIRR-seq resolves subisotype specific repertoire diversity.** (A) Boxplots of V
685 gene family usage frequencies in expressed IgG subisotypes across all ten individuals. (B) Bar
686 plots showing distribution of unique VDJ sequences across isotypes, subisotypes, and subisotype
687 alleles in two representative samples, 1008 and 602c, characterized by FLAIRR-seq. (C) Circos
688 plots showing V family gene usage frequency within each subisotype allele fraction for samples
689 1008 and 602c. (D) Table showing percentage of V gene family usage within each constant region
690 allele as represented in the circus plots in (C).

691 **Figure 5. FLAIRR-seq resolves subisotype-specific clonal expansion and evidence of CSR**
692 **in a patient hospitalized for COVID-19.** (A) Overview of experimental design: whole blood-
693 derived RNA was collected on days 1, 4, 8 and 13 post-hospitalization and used for FLAIRR-seq
694 profiling. (B) Bar plot showing percentage of unique clones per subisotype across all timepoints.
695 (C) Simpson's diversity index (q=2) for all clones in each subisotype across four timepoints; IgG4
696 not included due to low sequence number that did not pass filtering criteria. Decreased diversity
697 indicates increased clonal expansion. (D) Polarization, or number clones needed to comprise
698 80% of the repertoire, reported as percentage of total subisotype-specific repertoires across time.
699 Decreased polarity indicates increased clonal expansion. (E) Distribution of clone 13611 across
700 isotypes and subisotypes over time, suggesting CSR of this clone. (F) Phylogenetic tree of clone
701 13611 showing emergence of CSR transcript over time (red box).

702 References:

703 1. Schroeder, H. W., and L. Cavacini. 2010. Structure and function of immunoglobulins.
704 *Journal of Allergy and Clinical Immunology* 125: S41-S52.

705 2. Janda, A., A. Bowen, N. S. Greenspan, and A. Casadevall. 2016. Ig Constant Region
706 Effects on Variable Region Structure and Function. *Front Microbiol* 7: 22.

707 3. Nakano, T., M. Matsui, I. Inoue, T. Awata, S. Katayama, and T. Murakoshi. 2011. Free
708 immunoglobulin light chain: Its biology and implications in diseases. *Clinica Chimica Acta*
709 412: 843-849.

710 4. Lu, L. L., T. J. Suscovich, S. M. Fortune, and G. Alter. 2018. Beyond binding: antibody
711 effector functions in infectious diseases. *Nature Reviews Immunology* 18: 46-61.

712 5. Vidarsson, G., G. Dekkers, and T. Rispens. 2014. IgG subclasses and allotypes: from
713 structure to effector functions. *Front Immunol* 5: 520-520.

714 6. Greiff, V., C. R. Weber, J. Palme, U. Bodenhofer, E. Miho, U. Menzel, and S. T. Reddy.
715 2017. Learning the High-Dimensional Immunogenomic Features That Predict Public and
716 Private Antibody Repertoires. *The Journal of Immunology* 199: 2985.

717 7. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575-581.

718 8. Nishana, M., and S. C. Raghavan. 2012. Role of recombination activating genes in the
719 generation of antigen receptor diversity and beyond. *Immunology* 137: 271-281.

720 9. Tong, P., A. Granato, T. Zuo, N. Chaudhary, A. Zuiani, S. S. Han, R. Donthula, A.
721 Shrestha, D. Sen, J. M. Magee, M. P. Gallagher, C. E. van der Poel, M. C. Carroll, and D.
722 R. Wesemann. 2017. IgH isotype-specific B cell receptor expression influences B cell fate.
723 *Proc Natl Acad Sci U S A* 114: E8411-e8420.

724 10. Noviski, M., J. L. Mueller, A. Satterthwaite, L. A. Garrett-Sinha, F. Brombacher, and J.
725 Zikherman. 2018. IgM and IgD B cell receptors differentially respond to endogenous
726 antigens and control B cell fate. *eLife* 7: e35074.

727 11. Stavnezer, J., J. E. Guikema, and C. E. Schrader. 2008. Mechanism and regulation of
728 class switch recombination. *Annu Rev Immunol* 26: 261-292.
729 12. Karagiannis, P., A. E. Gilbert, D. H. Josephs, N. Ali, T. Dodev, L. Saul, I. Correa, L.
730 Roberts, E. Beddowes, A. Koers, C. Hobbs, S. Ferreira, J. L. Geh, C. Healy, M. Harries,
731 K. M. Acland, P. J. Blower, T. Mitchell, D. J. Fear, J. F. Spicer, K. E. Lacy, F. O. Nestle,
732 and S. N. Karagiannis. 2013. IgG4 subclass antibodies impair antitumor immunity in
733 melanoma. *J Clin Invest* 123: 1457-1474.
734 13. Chung, A. W., T. K. Ho, P. Hanson-Manful, S. Tritscheller, J. M. Raynes, A. L. Whitcombe,
735 M. L. Tay, R. McGregor, N. Lorenz, J. R. Oliver, J. K. Gurney, C. G. Print, N. J. Wilson,
736 W. J. Martin, D. A. Williamson, M. G. Baker, and N. J. Moreland. 2020. Systems
737 immunology reveals a linked IgG3-C4 response in patients with acute rheumatic fever.
738 *Immunol Cell Biol* 98: 12-21.
739 14. Vander Heiden, J. A., P. Stathopoulos, J. Q. Zhou, L. Chen, T. J. Gilbert, C. R. Bolen, R.
740 J. Barohn, M. M. Dimachkie, E. Ciafaloni, T. J. Broering, F. Vigneault, R. J. Nowak, S. H.
741 Kleinstein, and K. C. O'Connor. 2017. Dysregulation of B Cell Repertoire Formation in
742 Myasthenia Gravis Patients Revealed through Deep Sequencing. *J Immunol* 198: 1460-
743 1473.
744 15. Huijbers, M. G., L. A. Querol, E. H. Niks, J. J. Plomp, S. M. van der Maarel, F. Graus, J.
745 Dalmau, I. Illa, and J. J. Verschuuren. 2015. The expanding field of IgG4-mediated
746 neurological autoimmune disorders. *European Journal of Neurology* 22: 1151-1161.
747 16. Polonelli, L., J. Pontón, N. Elguezabal, M. D. Moragues, C. Casoli, E. Pilotti, P. Ronzi, A.
748 S. Dobroff, E. G. Rodrigues, M. A. Juliano, D. L. Maffei, W. Magliani, S. Conti, and L. R.
749 Travassos. 2008. Antibody complementarity-determining regions (CDRs) can display
750 differential antimicrobial, antiviral and antitumor activities. *PLoS One* 3: e2371.

751 17. Liu, H., W. Pan, C. Tang, Y. Tang, H. Wu, A. Yoshimura, Y. Deng, N. He, and S. Li. 2021.
752 The methods and advances of adaptive immune receptors repertoire sequencing.
753 *Theranostics* 11: 8945-8963.
754 18. Trück, J., A. Eugster, P. Barennes, C. M. Tipton, E. T. Luning Prak, D. Bagnara, C. Soto,
755 J. S. Sherkow, A. S. Payne, M.-P. Lefranc, A. Farmer, A. C. The, M. Bostick, and E.
756 Mariotti-Ferrandiz. 2021. Biological controls for standardization and interpretation of
757 adaptive immune receptor repertoire profiling. *eLife* 10: e66274.
758 19. Horns, F., C. Vollmers, D. Croote, S. F. Mackey, G. E. Swan, C. L. Dekker, M. M. Davis,
759 and S. R. Quake. 2016. Lineage tracing of human B cells reveals the in vivo landscape of
760 human antibody class switching. *eLife* 5: e16578.
761 20. Calonga-Solís, V., D. Malheiros, M. H. Beltrame, L. d. B. Vargas, R. M. Dourado, H. C.
762 Issler, R. Wassem, M. L. Petzl-Erler, and D. G. Augusto. 2019. Unveiling the Diversity of
763 Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian
764 Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural
765 Selection. *Front Immunol* 10.
766 21. Jonsson, S., G. Sveinbjornsson, A. L. de Lapuente Portilla, B. Swaminathan, R. Plomp,
767 G. Dekkers, R. Ajore, M. Ali, A. E. H. Bentlage, E. Elmér, G. I. Eyjolfsson, S. A.
768 Gudjonsson, U. Gullberg, A. Gylfason, B. V. Halldorsson, M. Hansson, H. Holm, Å.
769 Johansson, E. Johnsson, A. Jonasdottir, B. R. Ludviksson, A. Oddsson, I. Olafsson, S.
770 Olafsson, O. Sigurdardottir, A. Sigurdsson, L. Stefansdottir, G. Masson, P. Sulem, M.
771 Wuhrer, A.-K. Wihlborg, G. Thorleifsson, D. F. Gudbjartsson, U. Thorsteinsdottir, G.
772 Vidarsson, I. Jonsdottir, B. Nilsson, and K. Stefansson. 2017. Identification of sequence
773 variants influencing immunoglobulin levels. *Nature Genetics* 49: 1182-1191.
774 22. Buck, D., E. Albrecht, M. Aslam, A. Goris, N. Hauenstein, A. Jochim, S. Cepok, V.
775 Grummel, B. Dubois, A. Berthele, P. Lichtner, C. Gieger, J. Winkelmann, and B. Hemmer.

776 2013. Genetic variants in the immunoglobulin heavy chain locus are associated with the
777 IgG index in multiple sclerosis. *Ann Neurol* 73: 86-94.
778 23. van Erp, E. A., W. Luytjes, G. Ferwerda, and P. B. van Kasteren. 2019. Fc-Mediated
779 Antibody Effector Functions During Respiratory Syncytial Virus Infection and Disease.
780 *Front Immunol* 10.
781 24. Jefferis, R., J. Lund, and J. D. Pound. 1998. IgG-Fc-mediated effector functions: molecular
782 definition of interaction sites for effector ligands and the role of glycosylation.
783 *Immunological Reviews* 163: 59-76.
784 25. Charney, A. W., N. W. Simons, K. Mouskas, L. Lepow, E. Cheng, J. Le Berichel, C. Chang,
785 R. Marvin, D. M. Del Valle, S. Calorossi, A. Lansky, L. Walker, M. Patel, H. Xie, N. Yi, A.
786 Yu, G. Kang, A. Mendoza, L. E. Liharska, E. Moya, M. Hartnett, S. Hatem, L. Wilkins, M.
787 Eaton, H. Jamal, K. Tuballes, S. T. Chen, A. Tabachnikova, J. Chung, J. Harris, C.
788 Batchelor, J. Lacunza, M. Yishak, K. Argueta, N. Karekar, B. Lee, G. Kelly, D. Geanon, D.
789 Handler, J. Leech, H. Stefanos, T. Dawson, I. Scott, N. Francoeur, J. S. Johnson, A. Vaid,
790 B. S. Glicksberg, G. N. Nadkarni, E. E. Schadt, B. D. Gelb, A. Rahman, R. Sebra, G.
791 Martin, C. Agashe, P. Agrawal, A. Akyatan, K. Alesso-Carra, E. Alibo, K. Alvarez, A.
792 Amabile, S. Ascolillo, R. Bailey, P. Begani, P. B. Correra, S.-A. Brown, M. Buckup, L.
793 Burka, L. Cambron, G. Carrara, S. Chang, J. Chien, M. Chowdhury, C. C. Bozkus, P.
794 Comella, D. Cosgrove, F. Cossarini, L. Cotter, A. Dave, B. Dayal, M. Dhainaut, R.
795 Dornfeld, K. Dul, N. Eber, C. Elaiho, F. Fabris, J. Faith, D. Falci, S. Feng, B. Fennessy, M.
796 Fernandes, S. Gangadharan, J. Grabowska, G. Gyimesi, M. Hamdani, M. Herbinet, E.
797 Herrera, A. Hochman, G. E. Hoffman, J. Hook, L. Horta, E. Humblin, S. Karim, J. Kim, D.
798 Lebovitch, G. Lee, G. H. Lee, J. Lee, M. Leventhal, K. Lindblad, A. Livanos, R. Machado,
799 Z. Mahmood, K. Mar, S. Maskey, P. Matthews, K. Meckel, S. Mehandru, C. Mercedes, D.
800 Meyer, G. Mollaoglu, S. Morris, K. Nie, M. Nisenholtz, G. Ofori-Amanfo, K. Onel, M.
801 Ounadjela, V. Patel, C. Pruitt, S. Rathi, J. Redes, I. Reyes-Torres, A. Rodrigues, A.

802 Rodriguez, V. Roudko, E. Ruiz, P. Scalzo, P. Silva, A. S. Schanoski, M. Straw, S.
803 Tabachnikova, C. Teague, B. Upadhyaya, V. Van Der Heide, N. Vaninov, D. Wacker, H.
804 Walsh, C. M. Wilk, J. Wilson, K. M. Wilson, L. Xue, N.-a. Yeboah, S. Young, N. Zaks, R.
805 Zha, T. Marron, N. Beckmann, S. Kim-Schulze, S. Gnjatic, M. Merad, and C.-B. T. The
806 Mount Sinai. 2020. Sampling the host response to SARS-CoV-2 in hospitals under siege.
807 *Nature Medicine* 26: 1157-1158.
808 26. Woolley, C., J. Chariker, E. Rouchka, E. Ford, E. Hudson, S. Waigel, M. Smith, and T.
809 Mitchell. 2022. Reference long-read isoform-aware transcriptomes of four human
810 peripheral blood lymphocyte subsets
811 . *G3: Genes, Genomes, Genetics*
812 .
813 27. Gupta, N. T., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H.
814 Kleinstein. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin
815 repertoire sequencing data. *Bioinformatics* 31: 3356-3358.
816 28. Vander Heiden, J. A., G. Yaari, M. Uduman, J. N. H. Stern, K. C. O'Connor, D. A. Hafler,
817 F. Vigneault, and S. H. Kleinstein. 2014. pRESTO: a toolkit for processing high-throughput
818 sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 30: 1930-1932.
819 29. Nouri, N., and S. H. Kleinstein. 2018. A spectral clustering-based method for identifying
820 clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* 34: i341-
821 i349.
822 30. Nouri, N., and S. H. Kleinstein. 2020. Somatic hypermutation analysis for improved
823 identification of B cell clonal families from next-generation sequencing data. *PLOS*
824 *Computational Biology* 16: e1007977.
825 31. Hoehn, K. B., O. G. Pybus, and S. H. Kleinstein. 2022. Phylogenetic analysis of migration,
826 differentiation, and class switching in B cells. *PLOS Computational Biology* 18: e1009885.

827 32. Rodriguez, O. L., W. S. Gibson, T. Parks, M. Emery, J. Powell, M. Strahl, G. Deikus, K.
828 Auckland, E. E. Eichler, W. A. Marasco, R. Sebra, A. J. Sharp, M. L. Smith, A. Bashir, and
829 C. T. Watson. 2020. A Novel Framework for Characterizing Genomic Haplotype Diversity
830 in the Human Immunoglobulin Heavy Chain Locus. *Front Immunol* 11.
831 33. Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:
832 3094-3100.
833 34. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis,
834 R. Durbin, and G. P. D. P. Subgroup. 2009. The Sequence Alignment/Map format and
835 SAMtools. *Bioinformatics* 25: 2078-2079.
836 35. Martin, M., M. Patterson, S. Garg, S. O Fischer, N. Pisanti, G. W. Klau, A. Schöenhuth,
837 and T. Marschall. 2016. WhatsHap: fast and accurate read-based phasing. *bioRxiv*:
838 085050.
839 36. Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next-
840 generation sequencing data. *Bioinformatics (Oxford, England)* 28: 3150-3152.
841 37. Lefranc, M. P. 2001. IMGT, the international ImMunoGeneTics database. *Nucleic Acids*
842 *Res* 29: 207-209.
843 38. Al'Khafaji, A. M., J. T. Smith, K. V. Garimella, M. Babadi, M. Sade-Feldman, M. Gatzen,
844 S. Sarkizova, M. A. Schwartz, V. Popic, E. M. Blaum, A. Day, M. Costello, T. Bowers, S.
845 Gabriel, E. Banks, A. A. Philippakis, G. M. Boland, P. C. Blainey, and N. Hacohen. 2021.
846 High-throughput RNA isoform sequencing using programmable cDNA concatenation.
847 *bioRxiv*: 2021.2010.2001.462818.
848 39. Brochu, H. N., E. Tseng, E. Smith, M. J. Thomas, A. M. Jones, K. R. Diveley, L. Law, S.
849 G. Hansen, L. J. Picker, M. Gale, and X. Peng. 2020. Systematic Profiling of Full-Length
850 Ig and TCR Repertoire Diversity in Rhesus Macaque through Long Read Transcriptome
851 Sequencing. *The Journal of Immunology* 204: 3434.

852 40. Irvine, E. B., and G. Alter. 2020. Understanding the role of antibody glycosylation through
853 the lens of severe viral and bacterial diseases. *Glycobiology* 30: 241-253.
854 41. Alter, G., T. H. M. Ottenhoff, and S. A. Joosten. 2018. Antibody glycosylation in
855 inflammation, disease and vaccination. *Seminars in Immunology* 39: 102-110.
856 42. Collin, M. 2020. Antibody glycosylation as an immunological key in health and disease.
857 *Glycobiology* 30: 200-201.
858 43. Plomp, R., L. R. Ruhaak, H.-W. Uh, K. R. Reiding, M. Selman, J. J. Houwing-Duistermaat,
859 P. E. Slagboom, M. Beekman, and M. Wuhrer. 2017. Subclass-specific IgG glycosylation
860 is associated with markers of inflammation and metabolic health. *Scientific Reports* 7:
861 12325.
862 44. Trampert, D. C., L. M. Hubers, S. F. J. van de Graaf, and U. Beuers. 2018. On the role of
863 IgG4 in inflammatory conditions: lessons for IgG4-related disease. *Biochimica et*
864 *Biophysica Acta (BBA) - Molecular Basis of Disease* 1864: 1401-1409.

865

FIGURE 1.

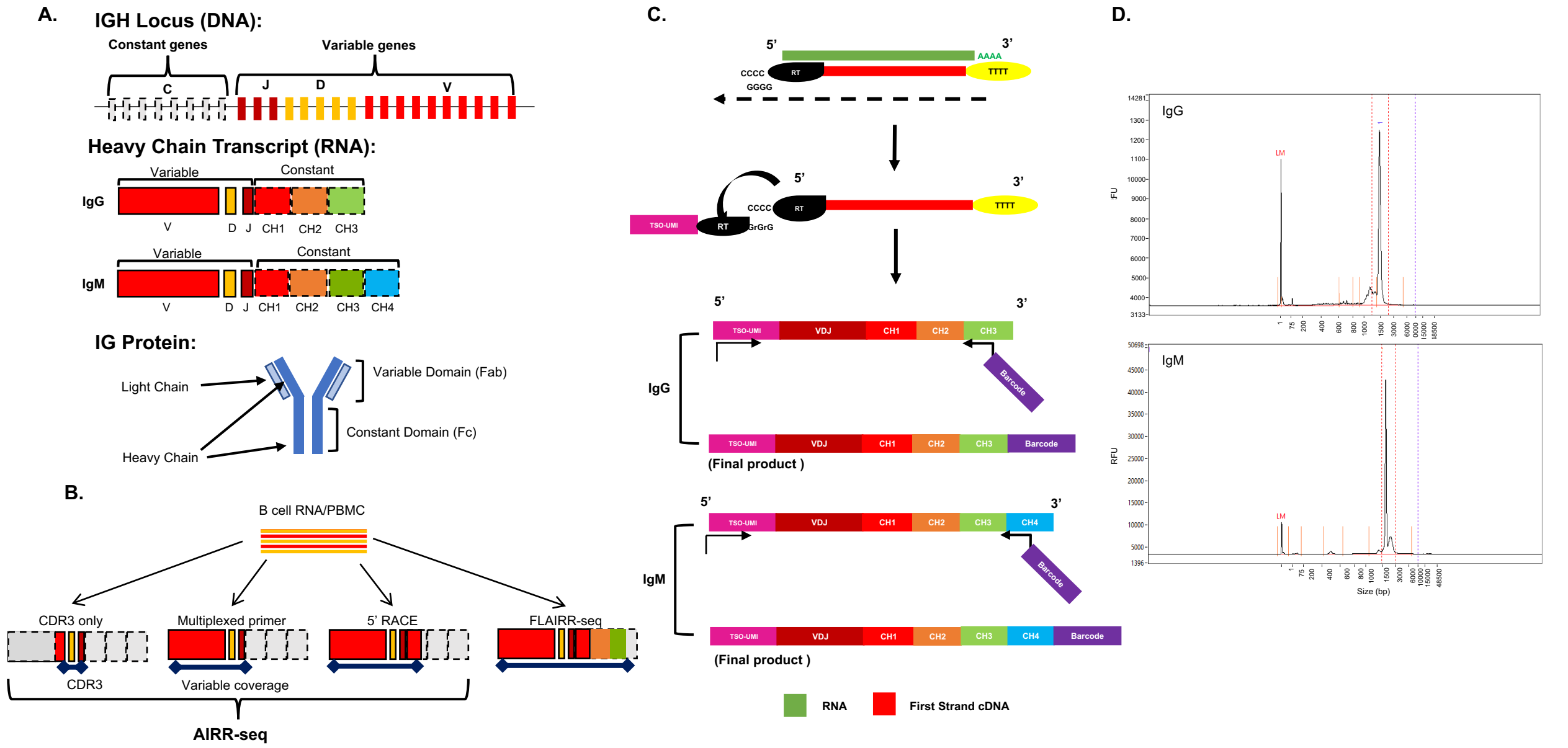
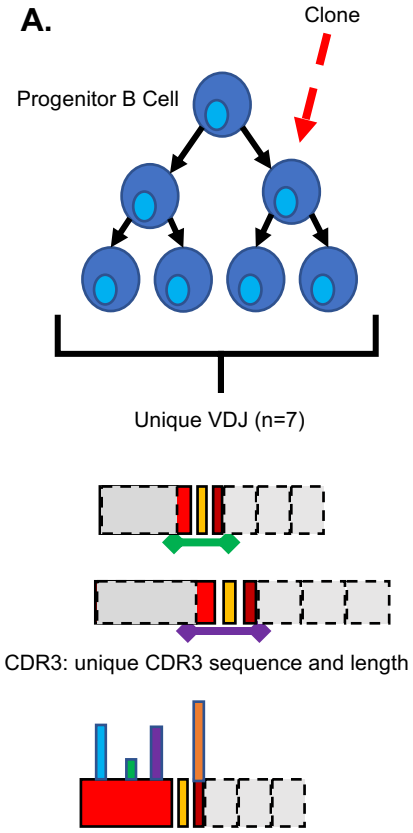

A.
IGH Locus (DNA):
Constant genes
Variable genes
C
J
D
V
Heavy Chain Transcript (RNA):
Variable
Constant
IgG
V
D
J
CH1
CH2
CH3
Variable
Constant
IgM
V
D
J
CH1
CH2
CH3
CH4
IG Protein:
Light Chain
Heavy Chain
Variable Domain (Fab)
Constant Domain (Fc)
B.
B cell RNA/PBMC
CDR3 only
Multiplexed primer
5' RACE
FLAIRR-seq
CDR3
Variable coverage
AIRR-seq
C.
5'
3'
AAAA
CCCC
RT
TTTT
GGGG
TSO-UMI
GrGrG
VDJ
CH1
CH2
CH3
CH4
Barcode
IgG
(Final product)
IgM
(Final product)
RNA
First Strand cDNA
D.
IgG
IgM
LM
RFU
Size (bp)

FIGURE 2.

A.



B. IgG BC vs. PBMC FLAIRR-seq

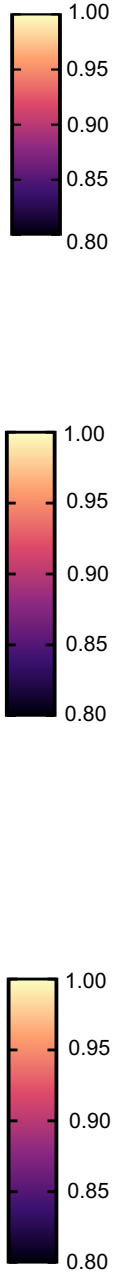
	V		D		J	
	r	p	r	p	r	p
1008	0.95	2.2e−16	0.95	2.2e−16	1.00	0.0028
0007	0.95	2.2e−16	0.95	4.1e−16	0.94	0.017
4002	0.95	2.2e−16	0.98	2.2e−16	0.89	0.033
705c	0.90	2.2e−16	0.97	2.2e−16	0.94	0.017

C. IgM

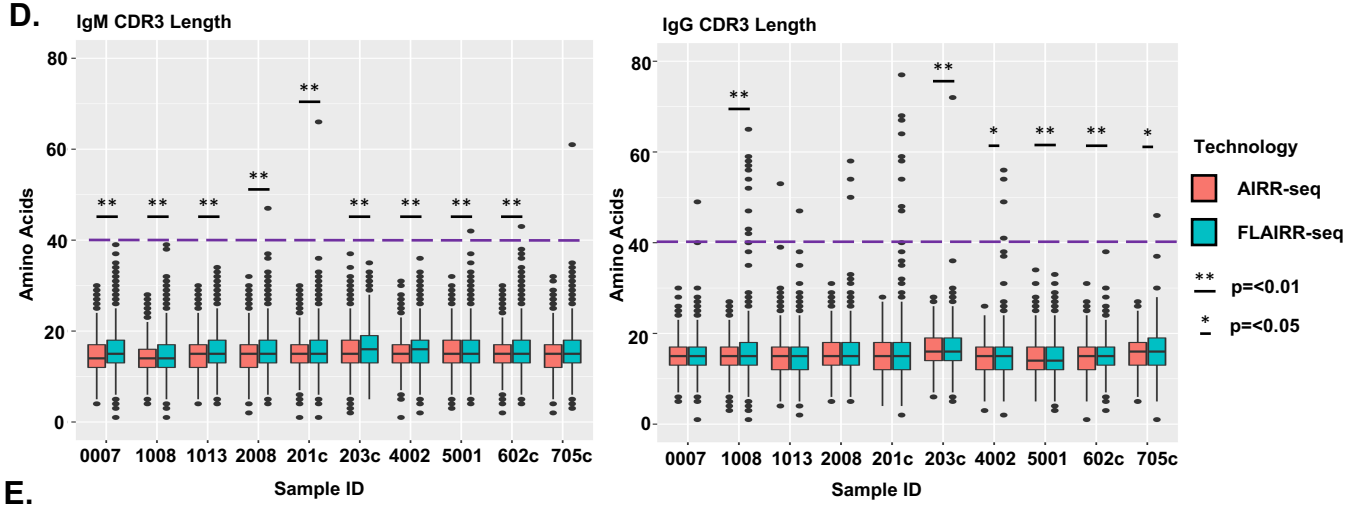
	V		D		J	
	r	p	r	p	r	p
0007	0.94	<2.2e-16	0.99	<2.2e-16	1.00	0.0028
201c	0.97	<2.2e-16	0.98	<2.2e-16	1.00	0.0028
203c	0.93	<2.2e-16	0.99	<2.2e-16	0.94	0.017
602c	0.96	<2.2e-16	0.99	<2.2e-16	1.00	0.0028
705c	0.95	<2.2e-16	0.99	<2.2e-16	0.94	0.017
1008	0.94	<2.2e-16	0.98	<2.2e-16	0.94	0.017
1013	0.96	<2.2e-16	0.99	<2.2e-16	0.94	0.017
2008	0.94	<2.2e-16	0.99	<2.2e-16	1.00	0.0028
4002	0.96	<2.2e-16	0.99	<2.2e-16	1.00	0.0028
5001	0.94	<2.2e-16	0.98	<2.2e-16	1.00	0.0028

IgG

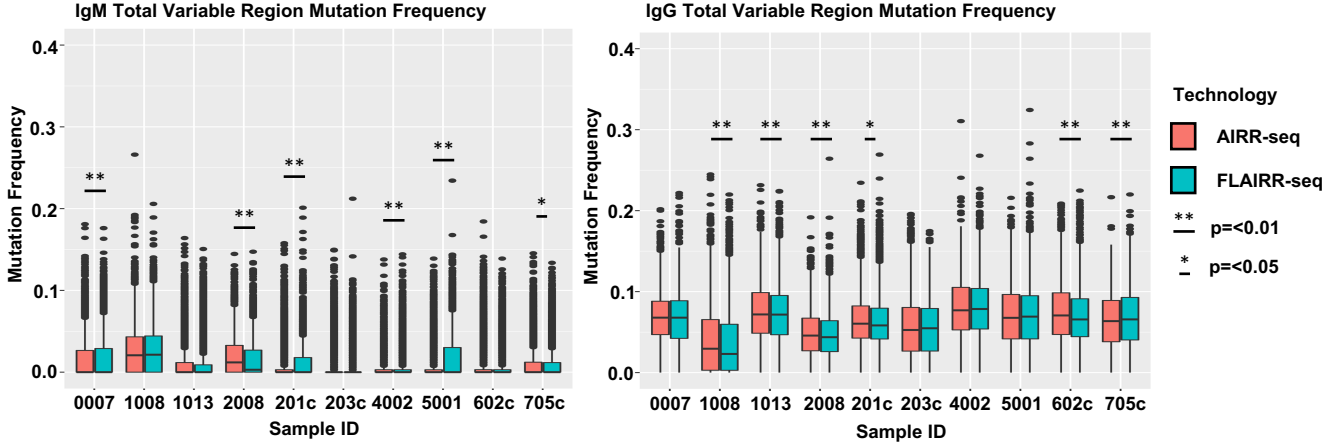
	V		D		J	
	r	p	r	p	r	p
0007	0.93	<2.2e-16	0.94	1.6e-15	0.94	0.017
201c	0.96	<2.2e-16	0.96	2.2e-16	0.89	0.033
203c	0.91	<2.2e-16	0.93	1.9e-14	0.94	0.017
602c	0.93	<2.2e-16	0.97	<2.2e-16	1.00	0.0028
705c	0.90	<2.2e-16	0.87	3.2e-10	1.00	0.0028
1008	0.95	<2.2e-16	0.99	<2.2e-16	1.00	0.0028
1013	0.95	<2.2e-16	0.95	2.2e-16	0.89	0.033
2008	0.95	<2.2e-16	0.97	2.2e-16	0.94	0.017
4002	0.95	<2.2e-16	0.93	6.1e-14	0.93	0.0077
5001	0.95	<2.2e-16	0.95	3.8e-16	1.00	0.0028



D.



E.



F. Iso-seq vs. FLAIRR-seq

	V		D		J	
	r	p	r	p	r	p
Sorted B cells	0.94	<2.2e-16	0.92	1.4e-13	1.00	0.0028

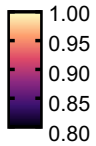


FIGURE 3.

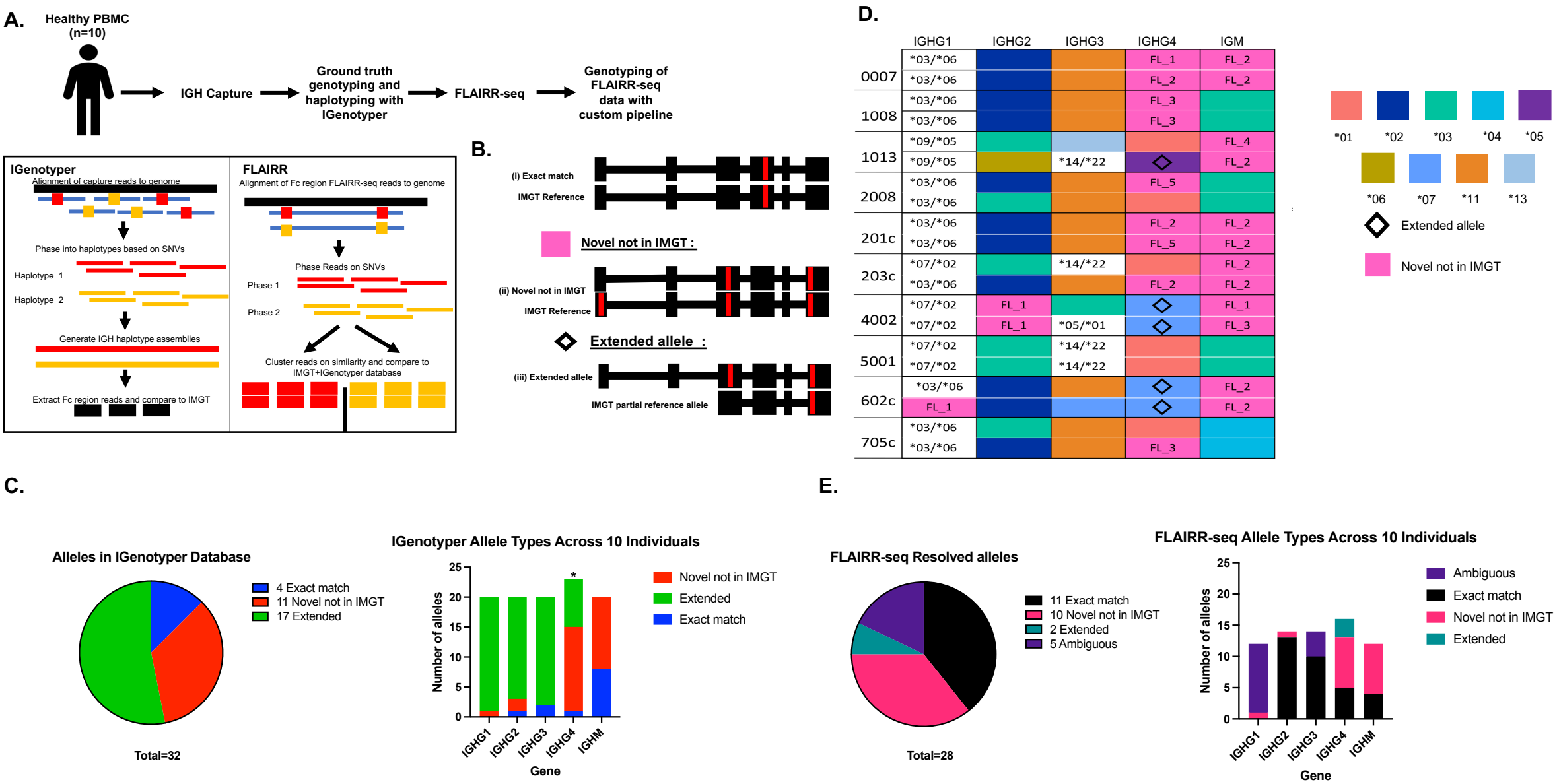
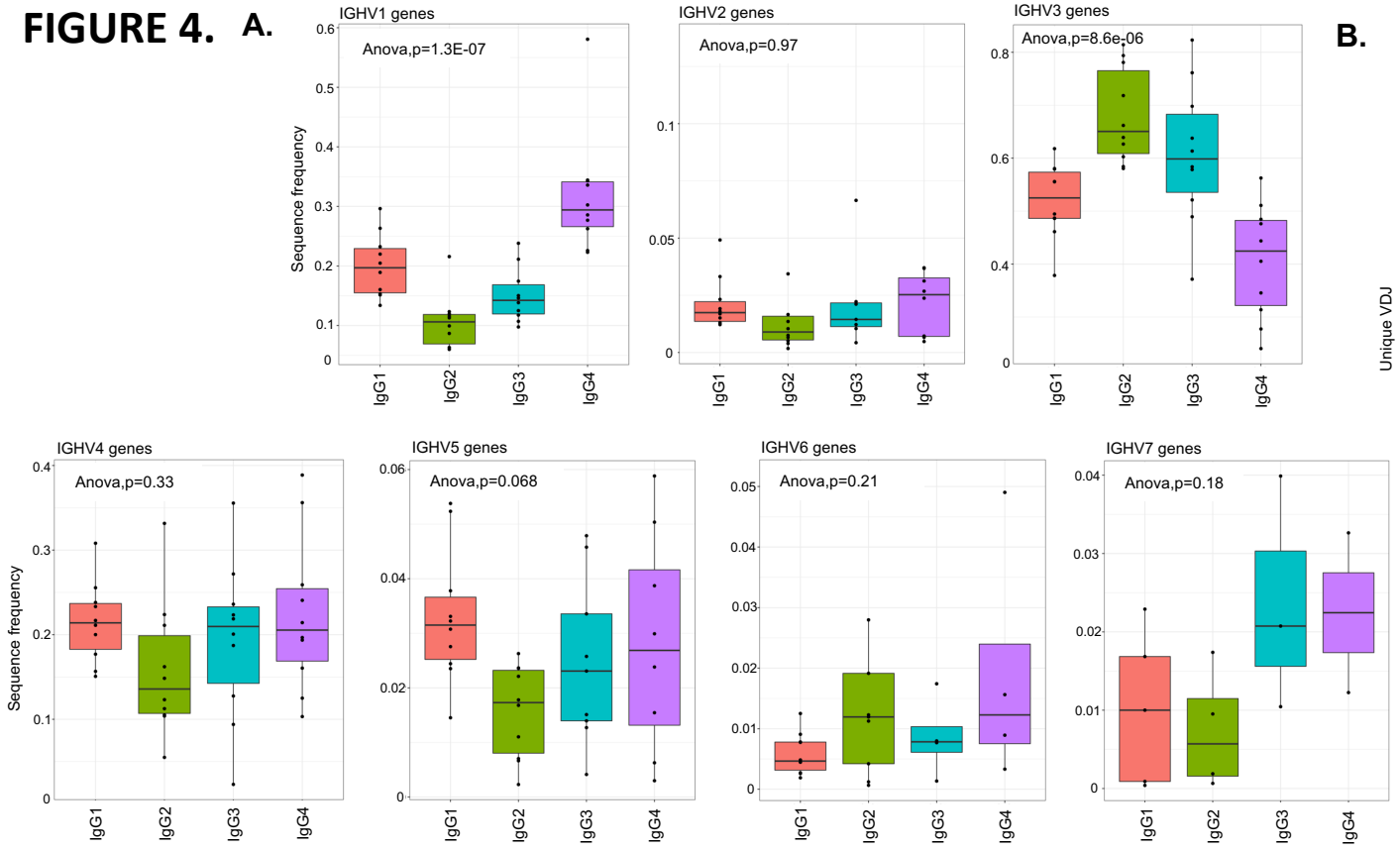
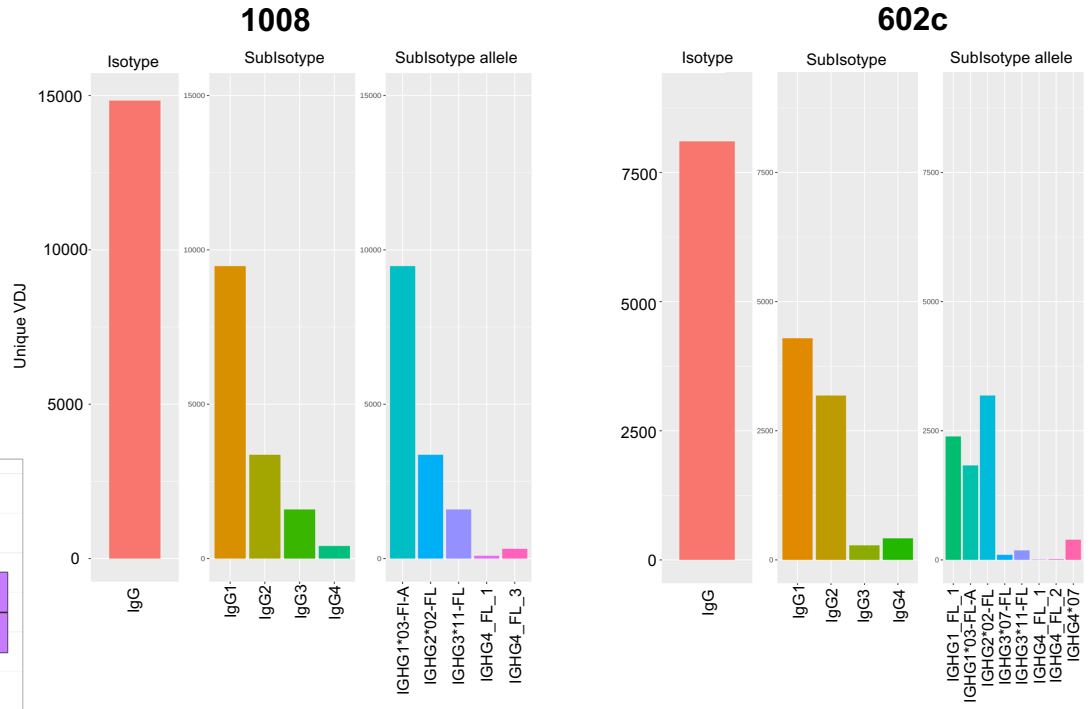

A.
Healthy PBMC (n=10)
IGH Capture
Ground truth genotyping and haplotyping with IGenotyper
FLAIRR-seq
Genotyping of FLAIRR-seq data with custom pipeline
IGenotyper
Alignment of capture reads to genome
Phase into haplotypes based on SNVs
Haplotype 1
Haplotype 2
Generate IGH haplotype assemblies
Extract Fc region reads and compare to IMGT
FLAIRR
Alignment of Fc region FLAIRR-seq reads to genome
Phase Reads on SNVs
Phase 1
Phase 2
Cluster reads on similarity and compare to IMGT+IGenotyper database
B.
(i) Exact match
IMGT Reference
Novel not in IMGT :
(ii) Novel not in IMGT
IMGT Reference
Extended allele :
(iii) Extended allele
IMGT partial reference allele
D.
IGHG1
IGHG2
IGHG3
IGHG4
IGM
0007
1008
1013
2008
201c
203c
4002
5001
602c
705c
*01
*02
*03
*04
*05
*06
*07
*11
*13
Extended allele
Novel not in IMGT
C.
Alleles in IGenotyper Database
4 Exact match
11 Novel not in IMGT
17 Extended
Total=32
IGenotyper Allele Types Across 10 Individuals
Number of alleles
Gene
IGHG1
IGHG2
IGHG3
IGHG4
IGHM
Novel not in IMGT
Extended
Exact match
E.
FLAIRR-seq Resolved alleles
11 Exact match
10 Novel not in IMGT
2 Extended
5 Ambiguous
Total=28
FLAIRR-seq Allele Types Across 10 Individuals
Number of alleles
Gene
Ambiguous
Exact match
Novel not in IMGT
Extended

FIGURE 4.

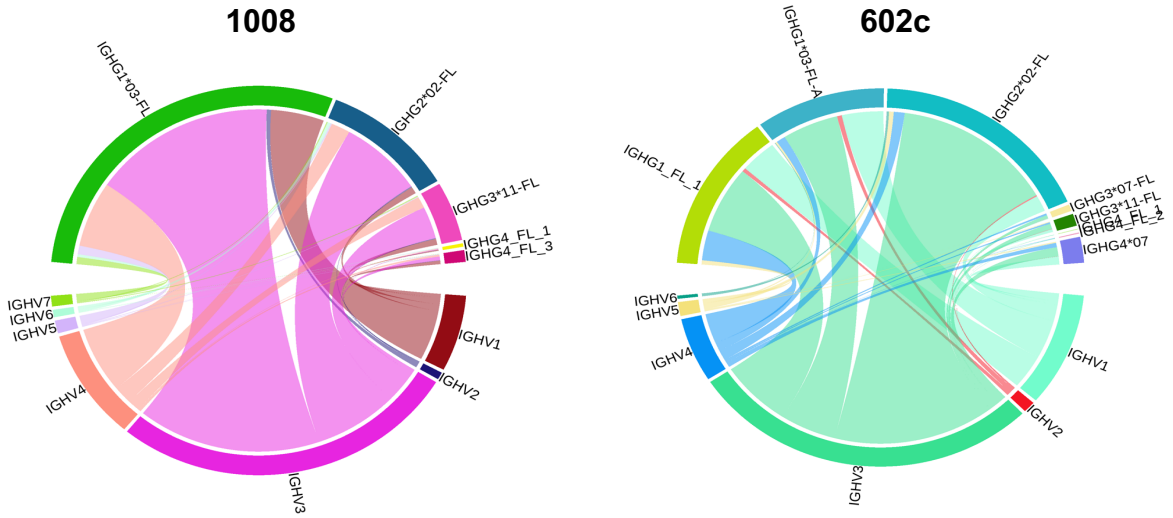
A.



B.



C.



D.

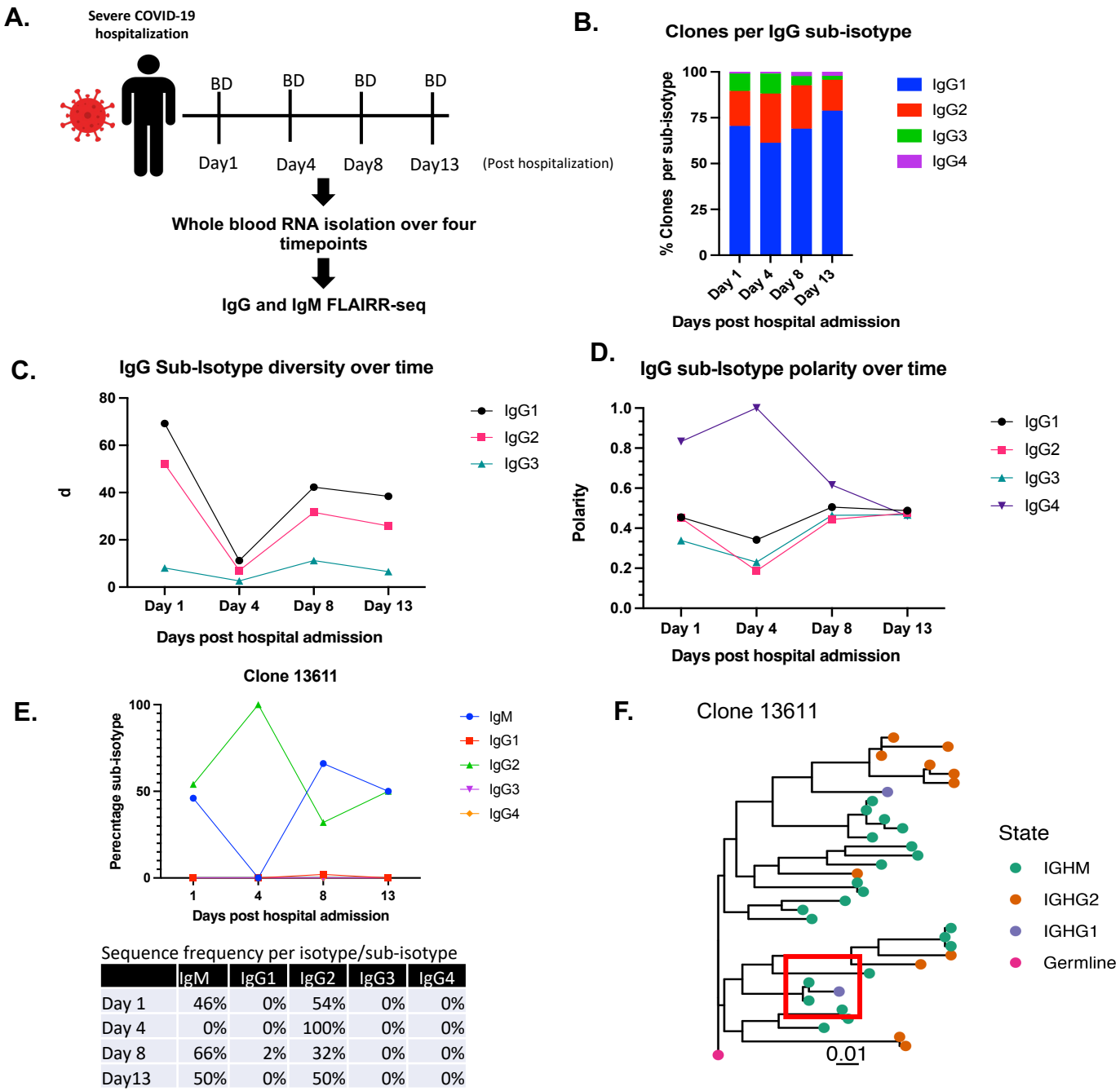
1008 V gene family usage

	IGHG1*03-FL	IGHG2*02-FL	IGHG3*11-FL	IGHG4_FL_1	IGHG4_FL_3
IGHV1	16%	6%	12%	33%	34%
IGHV2	1%	1%	1%	5%	3%
IGHV3	56%	72%	61%	15%	36%
IGHV4	21%	16%	22%	23%	18%
IGHV5	3%	1%	2%	10%	5%
IGHV6	1%	3%	0%	14%	2%
IGHV7	2%	1%	2%	0%	2%

602c V gene family usage

	IGHG1_FL_1	IGHG1*03-FL-A	IGHG2*02-FL	IGHG3*07-FL	IGHG3*11-FL	IGHG4_FL_1	IGHG4_FL_2	IGHG4*07
IGHGV1	25%	36%	9%	12%	26%	38%	19%	35%
IGHGV2	3%	4%	9%	2%	2%	0%	0%	1%
IGHGV3	48%	49%	9%	54%	51%	63%	38%	40%
IGHGV4	22%	8%	9%	31%	14%	0%	44%	19%
IGHGV5	3%	2%	9%	1%	7%	0%	0%	5%
IGHGV6	0%	1%	9%	0%	0%	0%	0%	0%
IGHGV7	0%	0%	9%	0%	0%	0%	0%	0%

FIGURE 5.



Sequence frequency per isotype/sub-isotype

	IgM	IgG1	IgG2	IgG3	IgG4
Day 1	46%	0%	54%	0%	0%
Day 4	0%	0%	100%	0%	0%
Day 8	66%	2%	32%	0%	0%
Day13	50%	0%	50%	0%	0%

Table 1: Comparative analysis metrics between AIRR-seq and FLAIRR-seq on matched samples

Sample ID	Isotype	Method	RNA input (ng)	Input reads into Immcantation[a]	Unique VDJ[b]	Unique Clones[c]	Unique CDR3[d]
0007	IgG	AIRR	100	**841627**	3164	1008	1187
		FLAIRR	335	181109	4880	1358	1520
201c	IgG	AIRR	100	**1509063**	4568	1221	1371
		FLAIRR	296	221301	10819	2145	2388
203c	IgG	AIRR	100	**1006965**	5793	1436	1705
		FLAIRR	106	475703	3778	993	1138
602c	IgG	AIRR	100	**1275902**	7736	1542	2217
		FLAIRR	149	344906	9225	1828	2357
705c	IgG	AIRR	100	**1113660**	1362	557	576
		FLAIRR	268	279601	2328	1063	1081
1008	IgG	AIRR	100	**1310350**	12217	3005	3307
		FLAIRR	335	238820	18235	3505	3813
1013	IgG	AIRR	100	**1158222**	5027	1795	1964
		FLAIRR	222	396322	7487	2565	2744
2008	IgG	AIRR	100	**835928**	3900	1235	1368
		FLAIRR	240	418906	4258	1308	1380
4002	IgG	AIRR	100	**632272**	1112	727	772
		FLAIRR	335	398838	5377	1719	1887
5001	IgG	AIRR	100	**695884**	1885	1134	1244
		FLAIRR	335	313843	12146	2226	2381
0007	IgM	AIRR	100	**841627**	9154	7924	8163
		FLAIRR	335	331380	14926	12367	12851
201c	IgM	AIRR	100	**1509063**	17888	14281	14622
		FLAIRR	296	148355	12803	9353	9573
203c	IgM	AIRR	100	**1006965**	18189	16295	16629
		FLAIRR	106	509536	10636	9315	9419
602c	IgM	AIRR	100	**1275902**	8378	7704	7833
		FLAIRR	149	273103	16406	14495	14745
705c	IgM	AIRR	100	**1113660**	14273	12936	13413
		FLAIRR	268	435391	17417	15706	16249
1008	IgM	AIRR	100	**1310350**	12710	7878	8368
		FLAIRR	335	181063	15284	8745	9276
1013	IgM	AIRR	100	**1158222**	9012	8057	8241
		FLAIRR	222	262617	17812	15291	15759
2008	IgM	AIRR	100	**835928**	8551	5838	6206
		FLAIRR	240	502354	14132	10225	10674
4002	IgM	AIRR	100	**632272**	11607	11184	11367
		FLAIRR	335	489287	22379	19000	21532
5001	IgM	FLAIRR	335	**313843**	12146	2226	2381
		FLAIRR	355	171688	6164	4866	5011

([a]) AIRR-seq data reflects total number of unassembled paired end reads (R1 and R2 bold) used as input into analyses.
([b]) Counts of unique sequences after Immcantation annotation
([c]) Counts of unique clones after clonal clustering
([d]) Counts of sequences containing unique CDR3 sequences