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], Ayelet Peres[4,5], 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], Gur Yaari[4,5], 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

[4] Faculty of Engineering, Bar Ilan University, 5290002 Ramat Gan, Israel

[5] Bar Ilan Institute of Nanotechnology and Advanced Materials, Bar Ilan University, 5290002, Ramat Gan, Israel

Corresponding Author:

Melissa L. Smith, PhD

ml.smith@louisville.edu

Abstract:

Current Adaptive Immune Receptor Repertoire Sequencing (AIRR-seq) strategies resolve expressed antibody (Ab) transcripts with limited resolution of the constant region. Here we present a novel near full-length AIRR-seq (FLAIRR-Seq) method that utilizes targeted amplification by 5' rapid amplification of cDNA ends (RACE), combined with single molecule, real-time sequencing to generate highly accurate (>Q40, 99.99%) IG heavy chain transcripts. FLAIRR-seq was benchmarked by comparing IG heavy chain variable (IGHV), diversity (IGHD), and joining (IGHJ) gene usage, complementarity-determining region 3 (CDR3) length, and somatic hypermutation to matched datasets generated with standard 5' RACE AIRR-seq and full-length isoform sequencing. 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 additionally resolving novel IG heavy chain constant (IGHC) gene features. FLAIRR-seq data provides, for the first time, simultaneous, single-molecule characterization of IGHV, IGHD, IGHJ, and IGHC region genes and alleles, allele-resolved subisotype definition, and high-resolution identification of class-switch recombination within a clonal lineage. In conjunction with genomic sequencing and genotyping of IGHC genes, FLAIRR-seq of the IgM and IgG repertoires from 10 individuals resulted in the identification of 32 unique IGHC alleles, 28 (87%) of which were previously uncharacterized. Together, these data demonstrate the capabilities of FLAIRR-seq to characterize IGHV, IGHD, IGHJ, and IGHC gene diversity for the most comprehensive view of bulk expressed Ab repertoires to date.

Introduction

Antibodies (Abs) or immunoglobulins (IGs) are the primary effectors of humoral immunity and are found as both membrane-bound receptors on B cells and circulating, secreted proteins (1). Both membrane-bound B cell receptors (BCRs) and secreted Abs act to recognize and bind antigen. All Abs and BCRs are composed of two identical heavy and light chains that are post-translationally associated. The heavy chain is comprised of two distinct domains: (i) the variable domain (Fab), which allows for antigen binding, and (ii) the constant domain (Fc), which modulates downstream effector functions (1, 2). The light chain also includes a variable domain that, once post-translationally associated with the heavy chain variable domain, interacts with cognate antigen (3). In humans, Abs are grouped into discrete isotypes and subisotypes (i.e., IgM, IgD, IgG1, IgG2, IgG3, IgG4, IgA1, IgA2, and IgE), based on the expression of specific constant (C) genes within the IG heavy chain locus (IGH). Each isotype and subisotype has unique effector properties that together represent the wide diversity of Ab-mediated functions, including binding of Fc receptors (FCR), activation of complement, opsonization, antibody-dependent cellular cytotoxicity (ADCC) and antibody-dependent cellular phagocytosis (ADCP) (4, 5).

To facilitate the development of diverse Ab repertoires capable of recognizing the wide range of pathogens humans encounter, the IG genomic loci are highly polymorphic and harbor diverse and complex sets of genes that recombine in each B cell to encode up to 10^{13} unique specificities (6). B cells create this expansive catalog of specificities through somatic recombination of the variable (V), diversity (D), and joining (J) genes in IGH, and V and J genes from the corresponding light chain loci, lambda (IGL) and kappa (IGK) (7). During VDJ recombination in IGH, a single D and J gene are first recombined, while the intervening and unselected D and J gene sequences are removed by RAG recombinase (8). After D and J genes are joined, further recombination of a specific V gene to the DJ gene cassette completes the formation of the full VDJ rearrangement. Following transcription of the recombined VDJ, a single

constant region gene is spliced together with the VDJ cassette to generate the completed heavy chain transcript (7). Recombination at IGL and IGK occurs similarly, recombining V and J genes only. Heavy and light chain transcripts are independently translated and linked via covalent cysteine bonds resulting in a fully functional protein prior to B cell cell-surface expression or secretion (Figure 1A) (9). Naïve B cells, which develop in the bone marrow from hematopoietic stem cell progenitors, have undergone VDJ recombination but have not yet encountered antigen, and solely express IgM and IgD (10). These naïve B cells then migrate to B cell zones in secondary lymphoid tissues where they encounter antigen, driving further maturation and class switch recombination (CSR) to enable the most effective humoral responses (11). CSR mediates the excision of IGHC genes at the DNA level, which leads to the utilization and linkage of different IGHC genes to the same VDJ, ultimately resulting in class switching to alternate isotypes and subisotypes (11).

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

Current Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) methods aim to resolve variable and constant region transcripts to differing extents. Profiling of the variable region, even in part, defines V, D and J gene usage while also providing characterization of complementarity determining regions (CDR) 1, 2, and 3, which are hypervariable and directly interact with target antigen (Figure 1B) (16). CDR3-targeted profiling approaches, such as those

104 used by Adaptive Biotechnologies (17), allow for V, D, and J gene assignments but do not provide
105 complete resolution of the entire variable region (17). Multiplexed primer-based AIRR-seq
106 strategies generate full variable region content but require specific primers to known targets and
107 therefore may miss novel genes and alleles. 5' RACE AIRR-seq methods capture the full-length
108 VDJ exon without variable region-targeted multiplexed primer pools therefore limiting the impact
109 of primer bias and enabling discovery of novel IGHV, IGHD, and IGHJ genes and alleles (18). 5'
110 RACE methods also often prime from the first IGHC exon (CH1), allowing for determination of
111 isotype. Additional methods have been developed that shift amplification strategies by capturing
112 additional IGHC region sequence to enable subisotype resolution; however, these methods
113 sacrifice full and contiguous characterization of the IGHV gene (19). All commonly used AIRR-
114 seq methods are further technically limited by the length restrictions (≤ 600nt) of short-read next
115 generation sequencing. As a result, no current AIRR-seq strategy resolves the complete heavy
116 chain transcript, including all IGHC exons alongside the recombined IGHV, IGHD, and IGHJ
117 genes. Our team has recently shown that population-based polymorphisms within the IGHV,
118 IGHD, IGHJ and IGL loci are far more extensive than previously known; the IGHC region has also
119 been shown to contain genomic diversity, although the extent of this diversity has likely not been
120 fully explored (20-27). Although it is understood that the Fc domain mediates Ab effector
121 functions, there is limited knowledge as to how genetic variation in this region may impact
122 functional capabilities or posttranslational modification (5, 28, 29). As such, there is a growing
123 need to understand genomic variation across the complete Ab molecule. To address these
124 limitations, we have developed novel, end-to-end pipeline to target, profile, and characterize the
125 Ab heavy chain repertoire in the context of isotype (IgG, IgM) and subisotype (IgG1, IgG2, IgG3,
126 IgG4).

127 Here, we present FLAIRR-seq, a novel targeted 5' RACE-based amplification of near full-
128 length IgG and IgM heavy chain transcripts, paired with single molecule real time (SMRT)
129 sequencing, resulting in highly accurate (mean read accuracy ~Q60, 99.9999%), near full-length

Ab sequences from RNA derived from whole blood, isolated PBMC, and purified B cells. When analyzed with the Immcantation AIRR-seq tool suite (30, 31), we demonstrate that FLAIRR-seq performs comparably to standard 5'RACE AIRR-seq methods and single-molecule isoform sequencing (Iso-Seq) strategies for characterizing the expressed Ab repertoire. We further highlight the novel features of FLAIRR-seq data, including phased identification of IGHV, IGHD, IGHJ, and IGHC genes, facilitating the profiling of subisotype- and IGHC allele-specific repertoires and CSR characterization.

Materials and Methods

Sample collections

Experiments were conducted using healthy donor peripheral blood mononuclear cells (PBMC), purified B cells from healthy donors, or whole blood collected from hospitalized COVID-19 patients (Supplementary Table 1). Commercially available healthy donor PBMC (STEMCELL Technologies) and a subset of matched purified B cells were utilized to generate AIRR-seq and FLAIRR-seq validation datasets. Full-length isoform sequencing (Iso-Seq) was performed using B cells isolated from the PBMC of a healthy, consented 57-year-old male donor at the University of Louisville (UofL) School of Medicine. The UofL Institutional Review Board approved sample collection (IRB 14.0661). For COVID-19 affected patient samples (n=5), whole blood was collected from the Mount Sinai COVID-19 biobank cohort of hospitalized COVID-19 patients, approved by the Institutional Review Board at the Icahn School of Medicine at Mount Sinai as previously described (32).

PBMC isolation and B cell purification

Frozen healthy donor PBMCs were purchased, thawed, and aliquoted for use in downstream experiments (STEMCELL Technologies). For Iso-Seq analyses, 175mL of venous blood was collected in a final concentration of 6mM K_3EDTA using standard phlebotomy. PBMCs

were isolated using Sepmate PBMC Isolation Tubes (STEMCELL Technologies) as previously described (33), with an additional granulocyte depletion step using the RosetteSep Human Granulocyte Depletion Cocktail (STEMCELL Technologies) as directed by the manufacturer. B cells from the freshly collected and frozen healthy donor PBMC were isolated using the EasySep Human Pan-B Cell Enrichment Kit, as described by the manufacturer (STEMCELL Technologies). Briefly, B cells, including plasma cells, were isolated by negative selection using coated magnetic particles. First, the B cell enrichment cocktail was added to the sample and mixed for a 5-minute incubation at room temperature, followed by addition of magnetic particles and further incubation for 5 minutes on the benchtop. The sample tube was then placed on an EasySep magnet (STEMCELL Technologies), and purified B cells were carefully eluted from the magnetic particles and immediately used for RNA extraction.

Genomic DNA and RNA extraction

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

FLAIRR-seq targeted amplification of heavy chain transcripts

Extracted RNA was thawed on ice and converted to first strand complementary DNA (cDNA) using the SMARTer RACE 5'/3' Kit (Takara Bio USA), as described by the manufacturer and a custom oligonucleotide that contained the template switch oligo and a unique molecular identifier (5' TSO-UMI) for template switch during first strand cDNA synthesis. The following reaction conditions were used: (i) a primary master mix was prepared with 4.0 µL 5X First-Strand Buffer, 0.5 µL DTT (100 mM), and 1.0 µL dNTPs (20 mM) per reaction and set aside until needed; (ii) in a separate 0.2-mL PCR tube, 10 µL of sample RNA and 1 µL 5'-CDS Primer A were combined and incubated in a thermal cycler at 72°C (lid temperature: 105°C) for 3 minutes, followed by cooling to 42°C for 2 minutes; (iii) after cooling, tubes were spun briefly to collect contents and 1µL (12µM) of the 5' TSO-UMI was added to the RNA; (iv) 0.5 µL of RNase inhibitor and 2.0 µL of SMARTScribe Reverse Transcriptase were added to the primary master mix tube per sample and 8 µL of the combined master mix was then added to each RNA-containing sample tube. First-strand cDNA synthesis reactions were incubated in a thermal cycler at 42°C (lid temperature: 105°C) for 90 mins, followed by heat inactivation at 70°C for 10 minutes. Total first strand cDNA generated in this reaction was diluted 1:2 with Tricine-EDTA Buffer before moving onto targeted heavy chain transcript amplification.

To specifically amplify heavy chain transcripts from total first-strand cDNA, targeted IgG and IgM transcript amplification reactions were performed using barcoded IgG (3' primer binding in the constant region exon 3, CH3) or IgM (3' primer binding in the constant region exon 4, CH4)-specific primers (Supplementary Table 2) and the following conditions: (i) 5 µL of diluted first-strand cDNA was added to 0.2-mL PCR tubes; (ii) a master mix was generated using 10 µL 5X PrimeSTAR GXL Buffer, 4 µL GXL dNTP mixture, 28 µL PCR-grade water, 1 µL PrimeSTAR GXL Polymerase and 1 µL 10x UPM form the SMARTer RACE 5'/3' Kit per reaction; (iii) 44 µL master mix was added to each reaction tube followed by 1 µL of the appropriate barcoded IgG (CH3) or IgM (CH4) primer. Different temperatures were used for annealing of IgG- (63.5°C) and IgM-

specific primers (60°C) to account for primer specific melting temperatures and to enhance targeted amplification specificity. Amplification conditions for full-length IgG were: 1 minute at 95°C, followed by 35 amplification cycles of 95°C for 30 sec, 63.5°C for 20 sec, and 2 minutes at 68°C, followed by a final extension for 3 minutes at 68°C and hold at 4°C. Amplification conditions for full-length IgM were: 1 minute at 95°C, followed by 35 amplification cycles of 95°C for 30 sec, 60°C for 20 sec., and 2 minutes at 68°C, followed by a final extension for 3 minutes at 68°C and hold at 4°C. Final amplification reactions were purified using a 1.1x (vol:vol) cleanup with ProNex magnetic beads (Promega) . Successfully amplified products were quantified with Qubit dsDNA HS assay (ThermoFisher Scientific) and length was evaluated with the Fragment Analyzer Genomic DNA HS assay (Agilent). Samples were equimolar pooled in 8-plexes for SMRTbell library preparation and sequencing.

FLAIRR-seq SMRTbell library preparation and sequencing

Eight-plex pools of targeted IgG or IgM amplicons were prepared into SMRTbell sequencing templates according to the “Procedure and Checklist for Iso-Seq Express Template for Sequel and Sequel II systems” protocol starting at the “DNA Damage Repair” step and using the SMRTbell Express Template Prep Kit 2.0, with some modifications (Pacific Biosciences). Briefly, targeted IgG and IgM amplicons underwent enzymatic DNA damage and end repair, followed by ligation with overhang SMRTbell adapters as specified in the protocol. To increase consistency in SMRTbell loading on the Sequel IIe system, we further treated the SMRTbell libraries with a nuclease cocktail to remove unligated amplified products, using the SMRTbell Enzyme Cleanup Kit, as recommended by the manufacturer (Pacific Biosciences). Briefly, after heat-killing the ligase with an incubation at 65°C, samples were treated with a nuclease cocktail at 37°C for 1 hour, and then purified with a 1.1X Pronex cleanup. Final SMRTbell libraries were evaluated for quantity and quality using the Qubit dsDNA HS assay and Fragment Analyzer Genomic DNA assay, respectively. Sequencing of each 8-plex, barcoded sample pool was

performed on one SMRTcell 8M using primer v4 and polymerase v2.1 on the Sequel IIe system with 30 hr movies. Demultiplexed, high-fidelity circular consensus sequence reads ("HiFi reads") were generated on the instrument for downstream analyses.

AIRR-seq SMARTer Human BCR IgG/IgM sequencing

Matched healthy donor RNA was used to generate targeted IgG and IgM AIRR-seq libraries using the SMARTer Human BCR IgG IgM H/K/L Profiling Kit (Takara Bio USA) according to the manufacturer's instructions with no modifications. Briefly, for each sample, proprietary IgG and IgM primers were used to amplify heavy chain transcripts following a 5'RACE reaction. AIRR-seq libraries were then quality controlled using the 2100 Bioanalyzer High Sensitivity DNA Assay Kit (Agilent) and the Qubit 3.0 Fluorometer dsDNA High Sensitivity Assay Kit. Sequencing on the MiSeq platform using 300 bp paired-end reads was performed using the 600-cycle MiSeq Reagent Kit v3 (Illumina) according to the manufacturer's instructions, and FASTQ reads were generated using the associated DRAGEN software package (Illumina).

B cell Iso-Seq

RNA extracted from healthy sorted B cells was used to generate Iso-Seq SMRTbell libraries following the "Procedure & Checklist Iso-Seq Express Template Preparation for the Sequel II System" with minor adaptations compared to the manufacturer's instructions. Briefly, Iso-Seq libraries were generated using 500 ng high-quality (RIN > 8) RNA as input into oligo-dT primed cDNA synthesis (NEB). Barcoded primers were incorporated into the cDNA during second strand synthesis. Following double-stranded cDNA amplification, transcripts from two samples sourced from purified B cells and NK cells were equimolar pooled as previously described (33). SMRTbells were generated from the pooled cDNA as described above for the FLAIRR-seq amplification products, including the addition of a nuclease digestion step. Quantity and quality of the final Iso-Seq libraries were performed with the Qubit dsDNA High Sensitivity Assay Kit and Agilent Fragment Analyzer Genomic DNA assay, respectively. This 2-plex Iso-Seq pool was

sequenced using primer v4 and polymerase v2.1 on the Sequel IIe system with a 30-hour movie. HiFi reads were generated on instrument before analyses. Demultiplexing of barcoded samples and generation of full-length non-concatemer (FLNC) predicted transcripts were performed using the Iso-Seq v3 pipeline available through SMRTLink (v.10.2). B-cell-derived FLNC reads were mapped to the human genome using the GMAP reference database and reads derived from chromosome 14 were extracted for downstream IGH transcript characterization via Immcantation, as described below.

Immcantation analyses of IgG and IgM repertoires

Analyses of FLAIRR-seq, AIRR-seq, and Iso-seq datasets were performed using Immcantation tools (30, 31). Demultiplexed barcoded HiFi (for SMRT sequencing data) or FASTQ (for AIRR-seq) reads were first processed using the pRESTO tool for quality control, UMI processing, and error profiling (30). For AIRR-seq, pRESTO analysis data from paired-end reads ("R1" and "R2") were trimmed to remove bases with < Q20 read quality and/or <125 bp length using the "FilterSeq trimqual" and "FilterSeq length", respectively. IgG and IgM CH3 or CH4 primer sequences were identified with an error rate of 0.2, and primers identified were then noted in FASTQ headers using "MaskPrimers align". Next, 12 basepair (bp) UMIs were located and extracted using "Maskprimers extract". Sequences found to have the same UMIs were grouped and aligned using "AlignSets muscle," with a consensus sequence generated for each UMI using "BuildConsensus". Mate pairing of AIRR-seq reads was conducted using a reference-guided alignment requiring a minimum of a 5 bp overlap via "AssemblePairs sequential". After collapsing consensus reads with the same UMI ("conscount") using "CollapseSeq"," reads with < 2 supporting sequences were removed from downstream analysis. For pRESTO processing of FLAIRR-seq, single HiFi reads ("R1") reads did not require trimming due to > Q20 sequence quality across all bases. 5'TSO-UMI and CH3 or CH4 region primers were identified along with a 22 bp UMI with an error rate of 0.3 using "MaskPrimers align". Reads were then grouped and

aligned using “AlignSets muscle”. Due to the single molecule nature of FLAIRR-seq reads, no mate pairing was required. Consensus reads were then generated as described above, including removal of sequences with < 2 supporting reads. Read counts following each step of data filtration for AIRR-seq and FLAIRR-seq are represented in Supplementary Tables 3 and 4, respectively.

pRESTO-filtered reads for both AIRR-seq and FLAIRR-seq data were then input into the Change-O tool (Table 1). Iso-seq required no initial processing from pRESTO and was input into Change-O for IG gene reference alignment along with AIRR-seq and FLAIRR-seq data using “igblastn”, clonal clustering using “DefineClones”, and germline reconstruction and conversion using “CreateGermlines.py” and the GRCh38 chromosome 14 germline reference (31). Fully processed and annotated data was then converted into a TSV format for use in downstream analyses. The Alakazam Immcantation tool suite was then used to quantify gene usage analysis, calculate CDR3 length, assess somatic hypermutation frequencies, and analyze clonal diversity (31). SCOPer clonal assignment by spectral clustering was conducted for COVID-19 patient time course samples (34, 35). For clonal lineage tree analysis, the Dowser tool was used to examine clonal diversity and CSR over time (36).

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

FLAIRR-seq validation samples also underwent IGHC targeted enrichment and long-read sequencing as previously described (37). Briefly, gDNA was mechanically sheared and size selected to include 5-9kb fragments using the BluePippin (Sage Science). Samples were then end repaired and A-tailed using the standard KAPA library preparation protocol (Roche). Universal priming sequences and barcodes were then ligated onto samples for multiplexing (Pacific Biosciences). Barcoded gDNA libraries were captured using IGH-specific probes following a SeqCap protocol (Roche). 26 IGH-enriched samples were purified and pooled together for SMRTbell library prep as described above, including the final nuclease digestion step. Pooled SMRTbells were annealed to primer v4, bound to polymerase v2.0, and sequenced on the Sequel

IIe system with 30h movies. After sequencing, HiFi reads were generated and analyzed by the IGenotyper pipeline (37). In brief, IGenotyper was used to detect single nucleotide variants and assemble sequences into haplotype-specific assemblies for downstream IGHC gene genotyping. Alleles were then extracted from assemblies using a bed file containing coordinates for each IGHC gene exon. After sequences were extracted, reads were then aligned to the IMGT database (downloaded on 2/21/22) and assigned as an exact match to IMGT, "novel" if there was no match to the IMGT database or "novel, extended" if a match was detected to a partial allele found in IMGT, but the IMGT allele was a substring of the IGenotyper identified allele (38). This set of alleles was then used as a ground truth dataset.

IGHC gene genotyping with FLAIRR-seq

To genotype IGHC genes and alleles from FLAIRR-seq data, productive reads were filtered by IGHC length (900bp–1100bp) and aligned to the chromosome 14 hg38 reference using minimap2 along with SAMtools to generate sorted and indexed bam files (39, 40). WhatsHap was used to identify, genotype, and phase single nucleotide variants (SNV) (41). Phased SNVs were used to assign each read to a haplotype using MsPAC. Reads from each haplotype and gene were clustered using CD-HIT using a 100% identity clustering threshold parameter, and a single representative read from the largest cluster was aligned to the IGenotyper curated alleles using BLAST (42) to determine the closest matching IGHC gene and allele (38, 43). The representative read was selected based on 100% identity to all other sequences in that cluster.

Inference of IGHV, IGHD, and IGHJ gene haplotypes from FLAIRR-seq data using IGHC gene anchors

To test the ability of IGHC genes to be used for the inference of IGHV, IGHD, and IGHJ haplotypes from FLAIRR-seq data, we chose one sample that was heterozygous for both IGHM and IGHJ6 (IGHJ6 is standardly used for AIRR-seq haplotype inference). For this sample, we employed TigGER (31, 44-46) to infer novel IGHV alleles, and generate sample-level IGHV genotypes using

a Bayesian approach. Rearranged sequences within the Change-O table were then reannotated taking into account sample genotype and detected novel alleles. Updated annotations were then used to infer haplotypes using RAbHIT version 0.2.4 (47). Both IGHJ6 and IGHM were used as anchor points for haplotyping, and the resulting haplotypes were compared.

Results

Gene usage, CDR3, and SHM profiles characterized from FLAIRR-seq data are comparable to AIRR-seq and Iso-Seq

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

To evaluate and validate the capabilities of FLAIRR-seq, matched FLAIRR-seq and AIRR-seq analyses were performed on ten healthy donor PBMC samples. FLAIRR-seq data were filtered from the initial HiFi reads (>Q20) to include only >Q40 reads. The average read quality of these filtered reads was >Q60 (99.9999%), with a pass filter rate ranging from 88%-93% of total reads. AIRR-seq FASTQ bases were trimmed to retain sequences with an average quality of Q20 (99%). These filtered reads were used as input into the Immcantation suite, specifically the pREST-O and Change-O tools, for IGHV, IGHD, and IGHJ gene assignment, and repertoire

feature analyses, including identification of clones, extent of somatic hypermutation, and evaluation of CDR3 lengths. As shown in **Table 1,** fewer overall FLAIRR-seq reads were used as input into the Immcantation analyses, after required filtration and read assembly steps (which were not needed for the high-quality single-molecule FLAIRR-seq reads). FLAIRR-seq resulted in comparable or, in many cases, increased number of unique VDJ sequences, clones, and CDR3 sequences identified compared to the matched AIRR-seq-derived samples in both the IgG and IgM repertoires. These basic sequencing and initial analysis metrics demonstrated that FLAIRR-seq produced high-quality variable region data for detailed Ab repertoire analyses and is amenable to analysis using existing AIRR-seq analysis tools. While input RNA mass used for FLAIRR-seq was often more than for AIRR-seq, ongoing and future optimization of the method is aimed at reducing RNA input. Furthermore, a comparative cost analysis was performed. To obtain the closest metric to a direct comparison, we calculated the cost per “actionable read”, defined as the read number per sample and per method after read and length quality filtration and assembly, but prior to cluster consensus, performed in the pRESTO pipeline. This method was used to represent the total unique single molecule or assembled templates captured by either the FLAIRR-seq or AIRR-seq methods, respectively, that passed all necessary quality control criteria for downstream annotation and analyses irrespective of biologic repertoire diversity or clonality. Lastly, to remove the impact of pooling differences, we used this “per actionable read” cost to calculate the cost for the generation of 15,000 “actionable reads” as our standard price. For AIRR-seq this cost was $25.50 per sample, whereas for FLAIRR-seq, this cost was $33.57 per sample. These costs reflect needed reagents and consumables only, assume instrumentation access and do not include labor. Future optimization for FLAIRR-seq will include integrating a multiplexed array sequencing (MAS) step to concatenate reads and enhance overall depth of sequencing and multiplexing capacity per pool, resulting in decreasing costs per sample (48).

While optimizing FLAIRR-seq sample preparation, we examined whether upstream isolation of B cells before FLAIRR-seq molecular preparation would enhance the ability to detect IGHV, IGHD, and IGHJ gene usage. To do this, aliquots of PBMC (n=4) were split into two groups for RNA extraction: (i) RNA derived from bulk PBMC, and (ii) RNA isolated from purified pan B-cells, followed by FLAIRR-seq preparation, SMRT sequencing, and Immcantation analysis of both groups. IGHV, IGHD, and IGHJ gene usage correlations between groups are shown in Figure 2A and demonstrate a significant association (p-values ranging from 0.033 to $4.1e^{-16}$) strongly supporting the conclusion that B cell isolation before RNA extraction was not necessary to achieve comparable gene usage metrics. The limited differences that were observed could be explained by template sampling differences between the two experiments. Due to the strong associations observed and the ease of processing PBMC in bulk, we moved forward with RNA derived directly from PBMC aliquots for the remainder of our analyses.

We established FLAIRR-seq performance by comparing its output to the commonly used 5' RACE AIRR-seq method. 5' RACE AIRR-seq was chosen as it provides resolution of the complete variable region and a small portion of IGHC, allowing for isotype differentiation. Following matched preparation of both FLAIRR-seq and AIRR-seq libraries from healthy donor PBMC samples (n=10), we compared multiple repertoire features to benchmark FLAIRR-seq performance. First, we evaluated IGHV, IGHD, and IGHJ gene usage frequencies. We observed significant correlations between FLAIRR-seq and AIRR-seq datasets in IGHV, IGHD and IGHJ gene usage for both IgM (V genes: r = 0.93-0.97, $p=<2.2e^{-16}$; D genes: r = 0.98-0.99, $p=<2.2e^{-16}$; J genes: r = 0.94-1.0, p = 0.0028-0.017) and IgG isotypes (V genes: r = 0.90-0.96, $p=<2.2e^{-16}$; D genes: r = 0.87-0.99, $p=<2.2e^{-16}$-$6.1e^{-14}$;J genes: r = 0.89-1.0, p = 0.0028-0.033) (Figure 2B), indicating that FLAIRR-seq comparably resolves IGHV, IGHD, and IGHJ gene usage profiles. To note, IGHJ genes showed lower levels of significance (larger p-values) across all comparisons due to the relatively few genes that make up the IGHJ gene family compared to the more diverse

IGHV and IGHD families. We next investigated the performance of both methods in terms of resolving somatic hypermutation (SHM) frequencies (Figure 2C), and complementarity determining region 3 (CDR3) lengths (Figure 2D), which are often used as measures of evaluating B cell affinity maturation. Although we did observe occasional statistically significant differences in the SHM frequency between AIRR-seq and FLAIRR-seq data using the same samples, these differences were not seen across all samples, suggesting that sample-to-sample variation may drive this observation rather than technology-based discrepancies. We found that CDR3 lengths were consistently longer in the FLAIRR-seq datasets for both the IgM and IgG isotypes in most donors. The characterization of unusually long CDR3 regions (> 40 nt) in the IgG sequences with FLAIRR-seq is likely due to the higher contiguity and quality afforded by the longer read lengths, which are less likely to be spanned by short-read 2x300 bp paired-end sequencing strategies. Together, these data demonstrate that FLAIRR-seq achieves comparable gene usage profiles, and improved resolution of long CDR3 sequences.

Others have recognized the power of long-read sequencing to resolve B cell repertoires using bulk Iso-Seq methods, allowing for the examination of full-length transcripts from isolated B cells (49). The Iso-Seq method captures full-length transcripts expressing a poly(A) tail without bias through oligo dT-based priming. The trade-offs of this approach are throughput, depth, and cost, as Iso-Seq processing generates a complete transcriptome per sample without enrichment of heavy chain sequences, which then need to be filtered out and analyzed, resulting in a considerable amount of non-repertoire data that is discarded. To investigate whether the untargeted transcriptome-wide Iso-Seq method would resolve a qualitatively different repertoire than FLAIRR-seq, which would have indicated FLAIRR-seq-driven primer bias, we performed matched Iso-Seq and FLAIRR-seq on purified B-cell derived RNA. IGHV, IGHD, and IGHJ gene usage frequencies were compared between Iso-Seq and FLAIRR-seq datasets (Spearman's rank correlation), revealing significant correlations between usage profiles (V genes: $r = 0.94$, $p=2.2e^{-}$

[16]; D genes: r = 0.92, p =1.4e^{-13}; J genes: r = 1.0, p = 0.0028; Figure 2E). These data strongly suggest that FLAIRR-seq has very limited to no primer-driven bias compared to whole transcriptome data. Collectively, this benchmarking dataset confirmed that FLAIRR-seq is comparable other state-of-the-art methods, providing robust characterization of commonly used repertoire metrics, with limited increases in per sample cost.

IGenotyper and FLAIRR-seq provide constant region gene allele identification and allow for haplotyping of variable genes

The novel value added by FLAIRR-seq is improved resolution of IGHC, including estimation of IGHC gene and allele usage, subisotype identification, and phasing of variable and constant regions for comprehensive repertoire analysis**.** To evaluate the capabilities and accuracy of IGHC gene and allele identification with FLAIRR-seq, we first established a ground truth dataset of IGHC alleles for all 10 samples by targeted sequencing of the germline IGH locus (Figure 3A) using IGenotyper, as previously described (37). IGHG1, IGHG2, IGHG3, IGHG4 and IGHM alleles called by IGenotyper (see Methods) were assigned to one of three categories, schematized in Figure 3B: (i) “exact match” - alleles documented in the IMGT database; (ii) “novel not in IMGT” – alleles not documented in the IMGT database; or (iii) “extended” – alleles that matched partial alleles in the IMGT database (i.e., those only spanning a subset of exons), but were extended by sequences in our dataset. IGenotyper identified a total of 32 unique IGHG1, IGHG2, IGHG3, IGHG4 and IGHM alleles across all individuals, as schematized in Figure 3C. Among these 32 alleles, only 4 were documented in IMGT, the remaining represented novel alleles (n=11) or extensions (n=17) of known alleles. In aggregate, we observed a greater number of IGHG4 alleles than for any of the IGHG genes. Among these alleles were 4 sequences represented by suspected duplications of IGHG4. In fact, we observed 3 IGHG4 gene alleles in 4/10 samples, indicating the presence of gene duplications in these donors; in all cases, these alleles were also identified in the FLAIRR-seq data (see below). Given the relatively small size of this proof-of-

454 concept healthy donor cohort, the identification of 28 (87%) novel or extended alleles underscores
455 the extensive polymorphism In this region and reflects the paucity of information regarding this
456 locus in existing immunogenomics databases.

457 We next used the iGenotyper-derived IGHC gene database as the ground-truth for
458 evaluating the capability of FLAIRR-seq to identify and resolve IGHC gene alleles. Based on our
459 analysis workflow for identifying IGHC alleles from FLAIRR-seq data, we resolved 19/32 (59%)
460 iGenotyper alleles at 100% identity; no additional false-positive alleles were identified. Of the
461 alleles that were not unambiguously resolved by our FLAIRR-seq pipeline, 8 had allele defining
462 single nucleotide variants (SNVs) 3' of the FLAIRR-seq primers. The rate of true-positive allele
463 calls using FLAIRR-seq across all 10 samples ranged from 5% for IGHG1 to 90% for IGHM
464 (Figures 3C and 3D). As a result, the IGHC genotypes inferred by FLAIRR-seq have some
465 limitations, but on the whole allow for much greater resolution of IGHC variation in the expressed
466 repertoire than currently used methods. Future iterations of FLAIRR-seq will include primer
467 optimization to facilitate better sequence coverage in the 3' regions of the IGHC alleles and
468 improve the direct genotyping capabilities of the FLAIRR-seq method.

469 Previous studies have demonstrated the use of IGHJ6 heterozygosity to infer haplotypes
470 of V and D genes from AIRR-seq data (46, 50). However, the frequency of IGHJ6 heterozygotes
471 in the population can vary. Therefore, we wanted to assess the utility of leveraging IGHC
472 polymorphism resolved by FLAIRR-seq for haplotyping IGHV alleles with the publicly available
473 tool RAbHIT(46, 50, 51). We selected a single donor ("1013") from our cohort that was
474 heterozygous for both IGHJ6 (*02 and *03) and IGHM (FL_2 and FL_4). Importantly, we were
475 able to associate each IGHJ6 allele to the respective IGHM allele from the corresponding
476 haplotype (Figure 3E). After defining germline IGHV alleles using TIgGER (31), we generated and
477 compared IGHV haplotype inferences using either IGHJ6 or IGHM alleles as anchor genes using
478 RAbHIT (47)(Figure 3E). Although some allele assignments were ambiguous ("unknown") using

both methods, we observed a strong consensus between haplotype inferences using the two anchor genes. For haplotype 1, represented by IGHJ6*03 and IGHM_FL_4, the IGHJ6*03-derived haplotype had 35 IGHV genes for which either an allele or deletion call was made. When using IGHM_FL_4, the same allele/deletion calls were made for 33 of these genes; in addition, using IGHM as the anchor gene, assignments were made for an additional 5 IGHV genes that had "unknown" designations using IGHJ6. Similarly, of the allele/deletion calls made for 36 IGHV genes on haplotype 2 assigned to IGHJ6*02, 33 of these gene had identical assignments to IGHM_FL_2. Together these results indicate that IGHC variants can be utilized for haplotype inference from repertoire data when commonly used IGHJ or IGHD genes are homozygous in individuals of interest.

FLAIRR-seq enables isotype-, subisotype-, and allele-specific repertoire analyses

IGHG and IGHM alleles identified in each sample were used to annotate reads in each respective repertoire. These assignments allowed for partitioning of the repertoire by isotype, subisotype and IGHC allele (Figure 4). To demonstrate this, we utilized the same representative sample ("1013") that was heterozygous for all IGHC genes. As shown in Figure 4A, IGHC gene assignments allow for subisotype and allele level frequencies to be estimated as a proportion of the overall IgG and IgM repertoires. In addition, detailed analyses of the repertoire can be conducted within each of these compartments. For example, Figure 4B shows the frequencies of IGHV gene subfamilies for each IGHG and IGHM allele identified in this sample. Using standard AIRR-seq analyses, we would not be able to identify allele-resolved V gene usage or enrichment within subisotype populations, which is important for linking subisotype functionality to particular antigen-specific VDJ clones.

Through the partitioning of repertoire sequences by subisotype and IGHC allele, we found that FLAIRR-seq also allowed for trends to be assessed in aggregate across donors. To demonstrate this, we further examined V gene family usage partitioned by IgG subisotype across

all 10 healthy donors. This analysis revealed expected patterns, in that IGHV1, IGHV3 and IGHV4 subfamily genes were dominant across the 4 subisotypes (Figure 4C). However, we did observe significant variation in subfamily proportions between subisotypes, associated with distinct profiles in specific subisotypes (Figure 4C). Specifically, the estimated frequencies of IGHV1 and IGHV3 were statistically different between subisotypes ($P<0.01$, ANOVA); IGHV1 usage was elevated in IGHG1 and IGHG4, whereas IGHV3 was elevated in IGHG2 and IGHG3. These analyses demonstrate the unique capability of FLAIRR-seq to examine variation in the expressed repertoire at the level of isotype, subisotype, and IGHC allele. As samples sizes increase, we expect that a multitude of additional repertoire features will become accessible to this kind of analysis leading to novel discoveries linking VDJ and IGHC genetic signatures.

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

We wanted to investigate the utility of FLAIRR-seq in clinical samples, particularly to observe changes in immune repertoires over time. Ab responses are highly dynamic, with specific Ab clones expanding upon activation by antigen. We were interested to know if class switch recombination could be captured by FLAIRR-seq, given the capability to identify clones with the subisotype resolved. We had the opportunity to evaluate FLAIRR-seq resolved repertoires over time in four samples collected from one individual over their >13-day hospitalization for severe COVID-19 disease. Blood draws were taken on days 1, 4, 8, and 13 post-hospitalization (Figure 5A) and analyzed with FLAIRR-seq across all time points. After initial FLAIRR-seq processing and analysis, we defined unique clones using SCOPer, which clusters sequences based on CDR3 similarity and mutations in IGHV and IGHJ genes (34). This analysis allowed for the estimation of subisotype-specific clone counts across the four timepoints examined. Overall, we observed IgG1 dominated the repertoire at all four time points, but the proportion of subisotypes fluctuated over time (Figure 5B). Specifically, the IGHG2 and IGHG3-specific repertoires expanded from day 1 to

529 day 4, but then contracted in overall frequency from day 8 to day 13 (Figure 5B). To assess clonal
530 diversity within each subisotype repertoire across time, we calculated the Simpson's diversity
531 index (q=2) using Alakazam (31). All subisotype-specific repertoires became less diverse from
532 day 1 to day 4, suggesting clonal expansion across the IGHG repertoire (Figure 5C). To note,
533 IGHG4 was not included in diversity calculations because IGHG4 would have required higher
534 sequencing depth to ascertain diversity, given the overall lower expression of IgG4 transcripts in
535 this individual. Subisotype-specific repertoire polarity was also assessed by calculating the
536 fraction of clones needed to represent 80% of the total repertoire (Figure 5D), with lower fractions
537 representing more polarized and clonally expended repertoires. Results of this analysis were
538 consistent with the diversity index, demonstrating an increase in clonal expansion (i.e., decreased
539 polarity) at day 4 across IGHG1, IGHG2, and IGHG3 compartments, which returned to baseline
540 at later timepoints.

541 CSR mediates the switching of Abs from one sub/isotype to another. This occurs through
542 the somatic recombination of IGHC genes, which brings the switched/selected IGHC genes
543 adjacent to the recombined IGHV, IGHD, and IGHJ segments, facilitating transcription. The
544 switching of isotypes and subisotypes can result in changes to associated effector functions of
545 the Ab while maintaining antigen-specific variable regions. Given the ability of FLAIRR-seq to
546 resolve clones with subisotype and IGHC allele resolution, as proof-of-concept, we sought to
547 assess whether FLAIRR-seq could allow for more detailed haplotype-level analysis of CSR
548 through the course of infection. To do this, we identified the largest clones in our dataset that were
549 both represented by multiple isotypes and found across timepoints. In total, using SCOper (34),
550 we identified 19 unique clonal lineages that met this criteria. We focused our detailed analysis on
551 one of the largest clones, "9900" (Figure 5E), comprising IGHM, IGHG1, and IGHG2 sequences.
552 On day 1 post-hospitalization, clone 9900 sequences were represented by both IGHM and
553 IGHG2. At day 4, IGHG2 was the only isotype observed, whereas again at day 8, both IGHM and

554 IGHG2 were observed, as well as IGHG1. To visualize CSR, we built a phylogeny using Dowser
555 (36). Highlighted in the red box on the phylogenetic tree shown in Figure 5F, we observe a single
556 subclade that is represented by IGHM, IGHG1, and IGHG2.

557 We were also able to resolve IGHC alleles from this individual, with the exception of
558 IGHG1 alleles which were ambiguous. Critically, both IGHG1 and IGHG2 were heterozygous
559 (Figure 5G). Through the assignment IGHG alleles to haplotypes within this individual using
560 heterozygous V genes (IGHV3-7 and IGHV3-48), we were able to determine that sequences from
561 clone 9900 (Figure 5F) utilized IGHG1 and IGHG2 alleles from the same haplotype, associated
562 with IGHV3-7*07 and IGHV3-48*03 (31). This observation offers direct characterization of CSR
563 events occurring on the same chromosome. When we looked across the remaining clones (n=8)
564 in this dataset that spanned time points and were represented by IGHG1 and IGHG2 subisotypes,
565 we were able to confirm that all of these clones used IGHG alleles from the same haplotype.

566 Together, these data provide demonstrative proof-of-concept evidence that FLAIRR-seq
567 profiling performs robustly on clinical samples, including RNA directly extracted from whole blood.
568 In addition, these data provide novel repertoire resolution extending what would have been
569 possible with standard AIRR-seq methods, including analysis of subisotype-specific repertories,
570 evaluation of clonal expansion, and characterization of CSR in the IgG and IgM repertoires.

571 **Discussion**

572 Here we present the development, validation, and application of FLAIRR-seq, a novel
573 method to resolve near full-length Ab transcripts from bulk PBMC-, isolated B cell- and whole
574 blood-derived total RNA. FLAIRR-seq enables highly accurate, simultaneous resolution of
575 variable and constant regions and suggests that IGHC polymorphism is far more extensive than
576 previously assumed. FLAIRR-seq performed equivalent to or with increased resolution compared
577 to existing standard 5'RACE AIRR-seq methods when resolving V, D, and J gene calls, CDR3
578 lengths, and SHM signatures, suggesting that our CH3/CH4 targeting strategies did not

579 compromise variable region characterization while simultaneously adding the capability to resolve
580 IGHC variation. Little to no primer bias was observed when compared to Ab repertoire profiling
581 from total mRNA Iso-seq methods. FLAIRR-seq provides the novel ability to use IGHC gene
582 usage to identify subisotypes and genotype heavy chain transcripts, linking these data back to
583 evaluate subisotype-specific repertoires, clonal expansion, and CSR. Underscoring the
584 underappreciated extent of IGHC variation, our profiling of a restricted cohort of only 10 individuals
585 from relatively homogenous backgrounds still identified 4 and 7 completely novel IGHC alleles in
586 IgM and IgG, respectively, and extended an additional 17 alleles beyond which had been available
587 in the IMGT database.

588 The unique capabilities of FLAIRR-seq will allow for novel examination of Ab repertoires,
589 including the characterization of variable gene usage and clonotype distribution within unique
590 subisotype subsets. This perspective has the potential to provide key insights into dynamic Ab
591 responses in diseases known to be mediated by subisotype-specific processes or have skewed
592 subisotype distribution as predictive markers of disease, including Myasthenia gravis (mediated
593 by pathogenic autoantibodies across subisotypes that give rise to varied disease pathologies),
594 Acute Rheumatic Fever (associated with elevated IgG3), and melanoma (skewing towards IgG4
595 in late-stage disease thought to be indicative of tolerogenic responses and poor prognosis)(12-
596 14). These subisotype-specific repertoire profiling approaches may be the first step toward
597 identification of unique clones that mediate disease pathogenicity or serve as high-resolution
598 biomarkers to disease progression, as well as open the door for potential functional experiments
599 on subisotype clones of interest, including examining the functional impact of the novel IGHC
600 alleles identified here. Expanded population-based FLAIRR-seq profiling and curation of novel
601 IGHC alleles, particularly in conjunction with IGenotyper targeted genomic assembly efforts in
602 IGH, will be a significant first step in defining the full extent of variation in a region too long
603 assumed to be relatively invariant.

The Fc region is known to be critical for modulating differential Ab effector functions. These differential functionalities are currently understood to be regulated by differential posttranslational modifications, such as variable glycosylation (52-55). Future FLAIRR-seq profiling will be a valuable tool to investigate how genomic variation across IGHC genes impacts residue usage and resultant Fc receptor binding, signaling potential, crosslinking, and potential for posttranslational modification, all of which would be expected to alter downstream effector functions, such as ADCC, ADCP, and complement fixation.

We further demonstrate that FLAIRR-seq can effectively examine clonal expansion and CSR in longitudinal samples, demonstrating the feasibility of using FLAIRR-seq to resolve Ab repertoire dynamics. This increased resolution will further our understanding of Ab repertoire evolution in the transition of acute to chronic disease states, many of which are associated with overall IgG subisotype distribution changes that are thought to reflect the inflammatory milieu (14, 56). One example is advanced melanoma, where late-stage disease is characterized by elevated IgG4 compared to IgG1, which is believed to reflect a more tolerizing, pro-tumor environment (12). FLAIRR-seq examination of these samples may identify specific repertoire distribution patterns that could act as biomarkers of disease progression. Moving forward it is critical to account for all variability within the Ab repertoire for the most comprehensive understanding of repertoire dynamics and the myriad factors impacting Ab effector function. Future efforts will expand FLAIRR-seq methods to target IGHA and IGHE repertoires, as well as implement multiplexed arrays (MAS) sequencing to considerably increase throughput and lower cost (48). Together, the data presented here demonstrate that the FLAIRR-seq method provides a comprehensive characterization of allele-resolved IgG and IgM repertoires, detailing variable region gene usage and measurements of maturation, isotype and subisotype identification, and the unappreciated extent of constant region variation, which will be necessary to fully appreciate the impact of IG genomic variation in health and disease.

Acknowledgements:

The authors would like to thank Kamille Rasche, Kaitlyin Shields, Uddalok Jana, and the staff of the University of Louisville Sequencing Technology Center for assistance in the sequencing and analysis of FLAIRR-seq samples. They would also like to acknowledge the support and assistance from Gur Yaari and Ayelet Peres for their assistance with the RabHIT workflow. This work was funded, in part, by P20 GM135004-02 (EEF, DT, CTW, and MLS).

References:

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

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

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

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

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

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

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

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

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

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

11. Stavnezer, J., J. E. Guikema, and C. E. Schrader. 2008. Mechanism and regulation of class switch recombination. *Annu Rev Immunol* 26: 261-292.

12. Karagiannis, P., A. E. Gilbert, D. H. Josephs, N. Ali, T. Dodev, L. Saul, I. Correa, L. Roberts, E. Beddowes, A. Koers, C. Hobbs, S. Ferreira, J. L. Geh, C. Healy, M. Harries, K. M. Acland, P. J. Blower, T. Mitchell, D. J. Fear, J. F. Spicer, K. E. Lacy, F. O. Nestle, and S. N. Karagiannis. 2013. IgG4 subclass antibodies impair antitumor immunity in melanoma. *J Clin Invest* 123: 1457-1474.

13. Chung, A. W., T. K. Ho, P. Hanson-Manful, S. Tritscheller, J. M. Raynes, A. L. Whitcombe, M. L. Tay, R. McGregor, N. Lorenz, J. R. Oliver, J. K. Gurney, C. G. Print, N. J. Wilson, W. J. Martin, D. A. Williamson, M. G. Baker, and N. J. Moreland. 2020. Systems immunology reveals a linked IgG3-C4 response in patients with acute rheumatic fever. *Immunol Cell Biol* 98: 12-21.

14. Vander Heiden, J. A., P. Stathopoulos, J. Q. Zhou, L. Chen, T. J. Gilbert, C. R. Bolen, R. J. Barohn, M. M. Dimachkie, E. Ciafaloni, T. J. Broering, F. Vigneault, R. J. Nowak, S. H. Kleinstein, and K. C. O'Connor. 2017. Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. *J Immunol* 198: 1460-1473.

15. Huijbers, M. G., L. A. Querol, E. H. Niks, J. J. Plomp, S. M. van der Maarel, F. Graus, J. Dalmau, I. Illa, and J. J. Verschuuren. 2015. The expanding field of IgG4-mediated neurological autoimmune disorders. *European Journal of Neurology* 22: 1151-1161.

16. Polonelli, L., J. Pontón, N. Elguezabal, M. D. Moragues, C. Casoli, E. Pilotti, P. Ronzi, A. S. Dobroff, E. G. Rodrigues, M. A. Juliano, D. L. Maffei, W. Magliani, S. Conti, and L. R. Travassos. 2008. Antibody complementarity-determining regions (CDRs) can display differential antimicrobial, antiviral and antitumor activities. *PLoS One* 3: e2371.

17. Liu, H., W. Pan, C. Tang, Y. Tang, H. Wu, A. Yoshimura, Y. Deng, N. He, and S. Li. 2021. The methods and advances of adaptive immune receptors repertoire sequencing. *Theranostics* 11: 8945-8963.

18. Trück, J., A. Eugster, P. Barennes, C. M. Tipton, E. T. Luning Prak, D. Bagnara, C. Soto, J. S. Sherkow, A. S. Payne, M.-P. Lefranc, A. Farmer, A. C. The, M. Bostick, and E. Mariotti-Ferrandiz. 2021. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. *eLife* 10: e66274.

19. Horns, F., C. Vollmers, D. Croote, S. F. Mackey, G. E. Swan, C. L. Dekker, M. M. Davis, and S. R. Quake. 2016. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *eLife* 5: e16578.

20. Calonga-Solís, V., D. Malheiros, M. H. Beltrame, L. d. B. Vargas, R. M. Dourado, H. C. Issler, R. Wassem, M. L. Petzl-Erler, and D. G. Augusto. 2019. Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection. *Front Immunol* 10.

21. Jonsson, S., G. Sveinbjornsson, A. L. de Lapuente Portilla, B. Swaminathan, R. Plomp, G. Dekkers, R. Ajore, M. Ali, A. E. H. Bentlage, E. Elmér, G. I. Eyjolfsson, S. A. Gudjonsson, U. Gullberg, A. Gylfason, B. V. Halldorsson, M. Hansson, H. Holm, Å. Johansson, E. Johnsson, A. Jonasdottir, B. R. Ludviksson, A. Oddsson, I. Olafsson, S. Olafsson, O. Sigurdardottir, A. Sigurdsson, L. Stefansdottir, G. Masson, P. Sulem, M. Wuhrer, A.-K. Wihlborg, G. Thorleifsson, D. F. Gudbjartsson, U. Thorsteinsdottir, G. Vidarsson, I. Jonsdottir, B. Nilsson, and K. Stefansson. 2017. Identification of sequence variants influencing immunoglobulin levels. *Nature Genetics* 49: 1182-1191.

22. Buck, D., E. Albrecht, M. Aslam, A. Goris, N. Hauenstein, A. Jochim, S. Cepok, V. Grummel, B. Dubois, A. Berthele, P. Lichtner, C. Gieger, J. Winkelmann, and B. Hemmer.

2013. Genetic variants in the immunoglobulin heavy chain locus are associated with the IgG index in multiple sclerosis. *Ann Neurol* 73: 86-94.

23. Keyeux, G., G. Lefranc, and M.-P. Lefranc. 1989. A multigene deletion in the human IGH constant region locus involves highly homologous hot spots of recombination. *Genomics* 5: 431-441.

24. Bashirova, A. A., W. Zheng, M. Akdag, D. G. Augusto, N. Vince, K. L. Dong, C. O'HUigin, and M. Carrington. 2021. Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. *Genes Immun* 22: 327-334.

25. Lefranc, M. P., G. Lefranc, G. de Lange, T. A. Out, P. J. van den Broek, J. van Nieuwkoop, J. Radl, A. N. Helal, H. Chaabani, E. van Loghem, and et al. 1983. Instability of the human immunoglobulin heavy chain constant region locus indicated by different inherited chromosomal deletions. *Mol Biol Med* 1: 207-217.

26. Lefranc, M.-P., and G. Lefranc. 2012. Human Gm, Km, and Am Allotypes and Their Molecular Characterization: A Remarkable Demonstration of Polymorphism. In *Immunogenetics: Methods and Applications in Clinical Practice*. F. T. Christiansen, and B. D. Tait, eds. Humana Press, Totowa, NJ. 635-680.

27. Lefranc, M.-P., G. Lefranc, and T. H. Rabbitts. 1982. Inherited deletion of immunoglobulin heavy chain constant region genes in normal human individuals. *Nature* 300: 760-762.

28. van Erp, E. A., W. Luytjes, G. Ferwerda, and P. B. van Kasteren. 2019. Fc-Mediated Antibody Effector Functions During Respiratory Syncytial Virus Infection and Disease. *Front Immunol* 10.

29. Jefferis, R., J. Lund, and J. D. Pound. 1998. IgG-Fc-mediated effector functions: molecular definition of interaction sites for effector ligands and the role of glycosylation. *Immunological Reviews* 163: 59-76.

734 30. Vander Heiden, J. A., G. Yaari, M. Uduman, J. N. H. Stern, K. C. O'Connor, D. A. Hafler,
735 F. Vigneault, and S. H. Kleinstein. 2014. pRESTO: a toolkit for processing high-throughput
736 sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 30: 1930-1932.
737 31. Gupta, N. T., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H.
738 Kleinstein. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin
739 repertoire sequencing data. *Bioinformatics* 31: 3356-3358.
740 32. Charney, A. W., N. W. Simons, K. Mouskas, L. Lepow, E. Cheng, J. Le Berichel, C. Chang,
741 R. Marvin, D. M. Del Valle, S. Calorossi, A. Lansky, L. Walker, M. Patel, H. Xie, N. Yi, A.
742 Yu, G. Kang, A. Mendoza, L. E. Liharska, E. Moya, M. Hartnett, S. Hatem, L. Wilkins, M.
743 Eaton, H. Jamal, K. Tuballes, S. T. Chen, A. Tabachnikova, J. Chung, J. Harris, C.
744 Batchelor, J. Lacunza, M. Yishak, K. Argueta, N. Karekar, B. Lee, G. Kelly, D. Geanon, D.
745 Handler, J. Leech, H. Stefanos, T. Dawson, I. Scott, N. Francoeur, J. S. Johnson, A. Vaid,
746 B. S. Glicksberg, G. N. Nadkarni, E. E. Schadt, B. D. Gelb, A. Rahman, R. Sebra, G.
747 Martin, C. Agashe, P. Agrawal, A. Akyatan, K. Alesso-Carra, E. Alibo, K. Alvarez, A.
748 Amabile, S. Ascolillo, R. Bailey, P. Begani, P. B. Correra, S.-A. Brown, M. Buckup, L.
749 Burka, L. Cambron, G. Carrara, S. Chang, J. Chien, M. Chowdhury, C. C. Bozkus, P.
750 Comella, D. Cosgrove, F. Cossarini, L. Cotter, A. Dave, B. Dayal, M. Dhainaut, R.
751 Dornfeld, K. Dul, N. Eber, C. Elaiho, F. Fabris, J. Faith, D. Falci, S. Feng, B. Fennessy, M.
752 Fernandes, S. Gangadharan, J. Grabowska, G. Gyimesi, M. Hamdani, M. Herbinet, E.
753 Herrera, A. Hochman, G. E. Hoffman, J. Hook, L. Horta, E. Humblin, S. Karim, J. Kim, D.
754 Lebovitch, G. Lee, G. H. Lee, J. Lee, M. Leventhal, K. Lindblad, A. Livanos, R. Machado,
755 Z. Mahmood, K. Mar, S. Maskey, P. Matthews, K. Meckel, S. Mehandru, C. Mercedes, D.
756 Meyer, G. Mollaoglu, S. Morris, K. Nie, M. Nisenholtz, G. Ofori-Amanfo, K. Onel, M.
757 Ounadjela, V. Patel, C. Pruitt, S. Rathi, J. Redes, I. Reyes-Torres, A. Rodrigues, A.
758 Rodriguez, V. Roudko, E. Ruiz, P. Scalzo, P. Silva, A. S. Schanoski, M. Straw, S.
759 Tabachnikova, C. Teague, B. Upadhyaya, V. Van Der Heide, N. Vaninov, D. Wacker, H.

760 Walsh, C. M. Wilk, J. Wilson, K. M. Wilson, L. Xue, N.-a. Yeboah, S. Young, N. Zaks, R.
761 Zha, T. Marron, N. Beckmann, S. Kim-Schulze, S. Gnjatic, M. Merad, and C.-B. T. The
762 Mount Sinai. 2020. Sampling the host response to SARS-CoV-2 in hospitals under siege.
763 *Nature Medicine* 26: 1157-1158.
764 33. Woolley, C., J. Chariker, E. Rouchka, E. Ford, E. Hudson, S. Waigel, M. Smith, and T.
765 Mitchell. 2022. Reference long-read isoform-aware transcriptomes of four human
766 peripheral blood lymphocyte subsets *G3: Genes, Genomes, Genetics*
767 34. Nouri, N., and S. H. Kleinstein. 2018. A spectral clustering-based method for identifying
768 clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* 34: i341-
769 i349.
770 35. Nouri, N., and S. H. Kleinstein. 2020. Somatic hypermutation analysis for improved
771 identification of B cell clonal families from next-generation sequencing data. *PLOS*
772 *Computational Biology* 16: e1007977.
773 36. Hoehn, K. B., O. G. Pybus, and S. H. Kleinstein. 2022. Phylogenetic analysis of migration,
774 differentiation, and class switching in B cells. *PLOS Computational Biology* 18: e1009885.
775 37. Rodriguez, O. L., W. S. Gibson, T. Parks, M. Emery, J. Powell, M. Strahl, G. Deikus, K.
776 Auckland, E. E. Eichler, W. A. Marasco, R. Sebra, A. J. Sharp, M. L. Smith, A. Bashir, and
777 C. T. Watson. 2020. A Novel Framework for Characterizing Genomic Haplotype Diversity
778 in the Human Immunoglobulin Heavy Chain Locus. *Front Immunol* 11.
779 38. Lefranc, M. P. 2001. IMGT, the international ImMunoGeneTics database. *Nucleic Acids*
780 *Res* 29: 207-209.
781 39. Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:
782 3094-3100.
783 40. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis,
784 R. Durbin, and G. P. D. P. Subgroup. 2009. The Sequence Alignment/Map format and
785 SAMtools. *Bioinformatics* 25: 2078-2079.

786 41. Martin, M., M. Patterson, S. Garg, S. O Fischer, N. Pisanti, G. W. Klau, A. Schöenhuth,
787 and T. Marschall. 2016. WhatsHap: fast and accurate read-based phasing. *bioRxiv*:
788 085050.
789 42. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local
790 alignment search tool. *J Mol Biol* 215: 403-410.
791 43. Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next-
792 generation sequencing data. *Bioinformatics (Oxford, England)* 28: 3150-3152.
793 44. Gadala-Maria, D., G. Yaari, M. Uduman, and S. H. Kleinstein. 2015. Automated analysis
794 of high-throughput B-cell sequencing data reveals a high frequency of novel
795 immunoglobulin V gene segment alleles. *Proceedings of the National Academy of*
796 *Sciences* 112: E862-E870.
797 45. Gadala-Maria, D., M. Gidoni, S. Marquez, J. A. Vander Heiden, J. T. Kos, C. T. Watson,
798 K. C. O'Connor, G. Yaari, and S. H. Kleinstein. 2019. Identification of Subject-Specific
799 Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. *Front Immunol* 10.
800 46. Gidoni, M., O. Snir, A. Peres, P. Polak, I. Lindeman, I. Mikocziova, V. K. Sarna, K. E. A.
801 Lundin, C. Clouser, F. Vigneault, A. M. Collins, L. M. Sollid, and G. Yaari. 2019. Mosaic
802 deletion patterns of the human antibody heavy chain gene locus shown by Bayesian
803 haplotyping. *Nature Communications* 10: 628.
804 47. Peres, A., M. Gidoni, P. Polak, and G. Yaari. 2019. RAbHIT: R Antibody Haplotype
805 Inference Tool. *Bioinformatics* 35: 4840-4842.
806 48. Al'Khafaji, A. M., J. T. Smith, K. V. Garimella, M. Babadi, M. Sade-Feldman, M. Gatzen,
807 S. Sarkizova, M. A. Schwartz, V. Popic, E. M. Blaum, A. Day, M. Costello, T. Bowers, S.
808 Gabriel, E. Banks, A. A. Philippakis, G. M. Boland, P. C. Blainey, and N. Hacohen. 2021.
809 High-throughput RNA isoform sequencing using programmable cDNA concatenation.
810 *bioRxiv*: 2021.2010.2001.462818.

811 49. Brochu, H. N., E. Tseng, E. Smith, M. J. Thomas, A. M. Jones, K. R. Diveley, L. Law, S.
812 G. Hansen, L. J. Picker, M. Gale, and X. Peng. 2020. Systematic Profiling of Full-Length
813 Ig and TCR Repertoire Diversity in Rhesus Macaque through Long Read Transcriptome
814 Sequencing. *The Journal of Immunology* 204: 3434.
815 50. Kidd, M. J., Z. Chen, Y. Wang, K. J. Jackson, L. Zhang, S. D. Boyd, A. Z. Fire, M. M.
816 Tanaka, B. A. Gaëta, and A. M. Collins. 2012. The Inference of Phased Haplotypes for
817 the Immunoglobulin H Chain V Region Gene Loci by Analysis of VDJ Gene
818 Rearrangements. *The Journal of Immunology* 188: 1333.
819 51. Kirik, U., L. Greiff, F. Levander, and M. Ohlin. 2017. Parallel antibody germline gene and
820 haplotype analyses support the validity of immunoglobulin germline gene inference and
821 discovery. *Mol Immunol* 87: 12-22.
822 52. Irvine, E. B., and G. Alter. 2020. Understanding the role of antibody glycosylation through
823 the lens of severe viral and bacterial diseases. *Glycobiology* 30: 241-253.
824 53. Alter, G., T. H. M. Ottenhoff, and S. A. Joosten. 2018. Antibody glycosylation in
825 inflammation, disease and vaccination. *Seminars in Immunology* 39: 102-110.
826 54. Collin, M. 2020. Antibody glycosylation as an immunological key in health and disease.
827 *Glycobiology* 30: 200-201.
828 55. Plomp, R., L. R. Ruhaak, H.-W. Uh, K. R. Reiding, M. Selman, J. J. Houwing-Duistermaat,
829 P. E. Slagboom, M. Beekman, and M. Wuhrer. 2017. Subclass-specific IgG glycosylation
830 is associated with markers of inflammation and metabolic health. *Scientific Reports* 7:
831 12325.
832 56. Trampert, D. C., L. M. Hubers, S. F. J. van de Graaf, and U. Beuers. 2018. On the role of
833 IgG4 in inflammatory conditions: lessons for IgG4-related disease. *Biochimica et*
834 *Biophysica Acta (BBA) - Molecular Basis of Disease* 1864: 1401-1409.
835
836

837 **Figure Legends:**

838 **Figure 1. Overview of Ab structure and FLAIRR-seq molecular method.** (A) Schematic
839 representation of the IGH locus, heavy chain transcript structure, and functional IG protein. (B)
840 Comparative coverage across the heavy chain transcript of commonly used AIRR-seq methods
841 compared with FLAIRR-seq. (C) FLAIRR-seq molecular pipeline: RNA (brown) was converted to
842 first-strand cDNA (red) using the 5' RACE method, incorporating a 5' TSO-UMI (pink) via template
843 switch. Second strand amplification specifically targeted IgG and IgM molecules through priming
844 of the 5' TSO-UMI and the 3' constant region IGH exon 3 (CH3) for IgG, or CH4 for IgM. A 16bp
845 barcode was incorporated into the 3' CH3/CH4 primers to enable sample multiplexing post-
846 amplification. (D) IGV screenshot showing near full-length single molecule structure of IGHG4
847 FLAIRR-seq transcripts.

848 **Figure 2. FLAIRR-seq shows robust characterization of V, D, and J genes.** (A) Spearman
849 ranked correlations and p-values of V, D, and J gene usage frequencies identified by FLAIRR-
850 seq performed on matched total PBMC and purified B cells. (B) Heatmap of Spearman ranked
851 correlations and p-values of V, D and J gene usage frequencies between FLAIRR-seq and AIRR-
852 seq processed samples. (C) Boxplots of somatic hypermutation frequencies defined by FLAIRR-
853 seq or AIRR-seq in IgM- and IgG-specific repertoires. (D) Boxplots of CDR3 length defined by
854 FLAIRR-seq or AIRR-seq analysis in both IgM- and IgG-specific repertoires. (E) Spearman
855 ranked correlation of V, D, and J gene usage frequencies between FLAIRR-seq-based and Iso-
856 Seq-based repertoire profiling. Significant differences between FLAIRR-seq and AIRR-seq data
857 indicated by * ($p < 0.05$) or ** ($p < 0.01$).

858 **Figure 3. FLAIRR-seq provides novel IGHC resolution for allelic discovery and allows**
859 **variable gene haplotyping.** (A) Overview of experimental design and pipeline overviews of
860 genotyping by IGenotyper (gDNA) and FLAIRR-seq (RNA). (B) Schematic depicting IGHC alleles
861 identified by IGenotyper, partitioned by identification as (i) exact matches to documented IMGT

alleles, (ii) novel alleles that are not in IMGT, or (iii) extended alleles. (C) Pie chart and stacked bar graph representing the total number of alleles and fraction of each category identified per IGHC gene as identified by either IGenotyper or FLAIRR-seq. Bar charts showing number of IGHC alleles from FLAIRR-seq that were resolved, ambiguous or unresolved when compared to IGenotyper alleles. * Indicates additional allele found due to IGHG4 duplication. (D) Table summarizing novel and extended alleles resolved by IGenotyper data. Extended alleles are denoted by *(allele number)-FL and novel alleles are denoted by FL_(number) alleles resolved by FLAIRR-seq are marked with a dot (•). (E) Venn diagrams showing number of IGHV haplotype allele/deletion calls when using IGHJ6 or IGHM anchors for each haplotype. Tile plots showing IGHV gene haplotypes inferred using either IGHJ6 anchors or IGHM anchors for one sample. Dark gray represents a deletion (DEL), off-white a non-reliable allele annotation (NRA), and light gray represents an unknown allele (Unk). Non-reliable alleles are annotated with an asterisk (*).

Figure 4. FLAIRR-seq resolves subisotype specific repertoire diversity. (A) Bar plots showing distribution of unique VDJ sequences across isotypes, subisotypes, and subisotype alleles in one representative sample, 1013, characterized by FLAIRR-seq. (B) Circos plots showing V family gene usage frequency within each subisotype allele for sample 1013. (C) Boxplots of V gene family usage frequencies within IGHG1, IGHG2, IGHG3, and IGHG4 repertoires across all ten individuals. (D) Principal component analysis of V gene family usage by subisotype; plot includes the first two principal components, and individual repertoires are colored by IGHG subisotype. (E) Boxplots showing sequence frequency of IGHV1 and IGHV3 family genes by subisotype across all 10 samples.

Figure 5. FLAIRR-seq resolves subisotype-specific clonal expansion and facilitates haplotype analysis of CSR in a patient hospitalized for COVID-19. (A) Overview of experimental design: whole blood-derived RNA was collected on days 1, 4, 8 and 13 post-hospitalization and used for FLAIRR-seq profiling. (B) Bar plot showing the percentage of clones

887 represented by each subisotype across timepoints. (C) Simpson's diversity index (q=2) for all
888 clones in each subisotype across four timepoints; IgG4 not included due to low sequence counts.
889 (D) Polarity, or the fraction of clones needed to comprise 80% of the repertoire, reported as
890 fraction of total subisotype-specific repertoires across time. (E) Distribution of a single clone
891 "9900" across isotypes and subisotypes over time, suggesting CSR of this clone. (F) Phylogenetic
892 tree constructed from sequences/members of clone 9900, with the inferred germline sequence
893 as the outgroup (star). Shapes and colors of tips (sequences) indicate time point and
894 isotype/subisotype. The scale bar represents the number of mutations between each node in the
895 tree. The subclade within the red box is represented by multiple time points and subisotypes,
896 providing evidence of CSR. (G) Tile plot showing the assignment of IGHC alleles to their
897 respective haplotypes, based on the frequency of observations in which each IGHC allele was
898 linked to each respective allele of heterozygous V genes, IGHV3-7 and IGHV3-48; light gray
899 denotes IGHC alleles for which haplotype assignment was not possible. Analysis of sequences
900 in (F) revealed that the IGHG1 and IGHG2 alleles represented in the phylogeny came from the
901 same haplotype (IGHG1*02/*07, IGHG2*08, IGHM_FL_2).
902

903

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]	Post assembly /Single reads[b]	Unique VDJc	Unique Clones[d]	Unique CDR3[c]
0007	IgG	AIRR	100	**841627**	15217	3164	1008	1187
		FLAIRR	335	181109	17881	4880	1358	1520
201c	IgG	AIRR	100	**1509063**	25015	4568	1221	1371
		FLAIRR	296	221301	36053	10819	2145	2388
203c	IgG	AIRR	100	**1006965**	23085	5793	1436	1705
		FLAIRR	106	475703	14923	3778	993	1138
602c	IgG	AIRR	100	**1275902**	38078	7736	1542	2217
		FLAIRR	149	344906	33903	9225	1828	2357
705c	IgG	AIRR	100	**1113660**	15358	1362	557	576
		FLAIRR	268	279601	18627	2328	1063	1081
1008	IgG	AIRR	100	**1310350**	56848	12217	3005	3307
		FLAIRR	335	238820	57020	18235	3505	3813
1013	IgG	AIRR	100	**1158222**	25250	5027	1795	1964
		FLAIRR	222	396322	27871	7487	2565	2744
2008	IgG	AIRR	100	**835928**	21681	3900	1235	1368
		FLAIRR	240	418906	16441	4258	1308	1380
4002	IgG	AIRR	100	**632272**	8949	1112	727	772
		FLAIRR	335	398838	19692	5377	1719	1887
5001	IgG	AIRR	100	**695884**	13930	1885	1134	1244
		FLAIRR	335	313843	42845	12146	2226	2381
0007	IgM	AIRR	100	**841627**	12777	9154	7924	8163
		FLAIRR	335	331380	46384	14926	12367	12851
201c	IgM	AIRR	100	**1509063**	26850	17888	14281	14622
		FLAIRR	296	148355	98069	12803	9353	9573
203c	IgM	AIRR	100	**1006965**	24623	18189	16295	16629
		FLAIRR	106	509536	37527	10636	9315	9419
602c	IgM	AIRR	100	**1275902**	12306	8378	7704	7833
		FLAIRR	149	273103	48897	16406	14495	14745
705c	IgM	AIRR	100	**1113660**	19484	14273	12936	13413
		FLAIRR	268	435391	79251	17417	15706	16249
1008	IgM	AIRR	100	**1310350**	20264	12710	7878	8368
		FLAIRR	335	181063	47531	15284	8745	9276
1013	IgM	AIRR	100	**1158222**	12769	9012	8057	8241
		FLAIRR	222	262617	52508	17812	15291	15759
2008	IgM	AIRR	100	**835928**	11608	8551	5838	6206
		FLAIRR	240	502354	47596	14132	10225	10674
4002	IgM	AIRR	100	**632272**	18411	11607	11184	11367
		FLAIRR	335	489287	91732	22379	19000	21532
5001	IgM	FLAIRR	335	**313843**	10018	12146	2226	2381
		FLAIRR	355	171688	18944	6164	4866	5011

904

FIGURE 1.

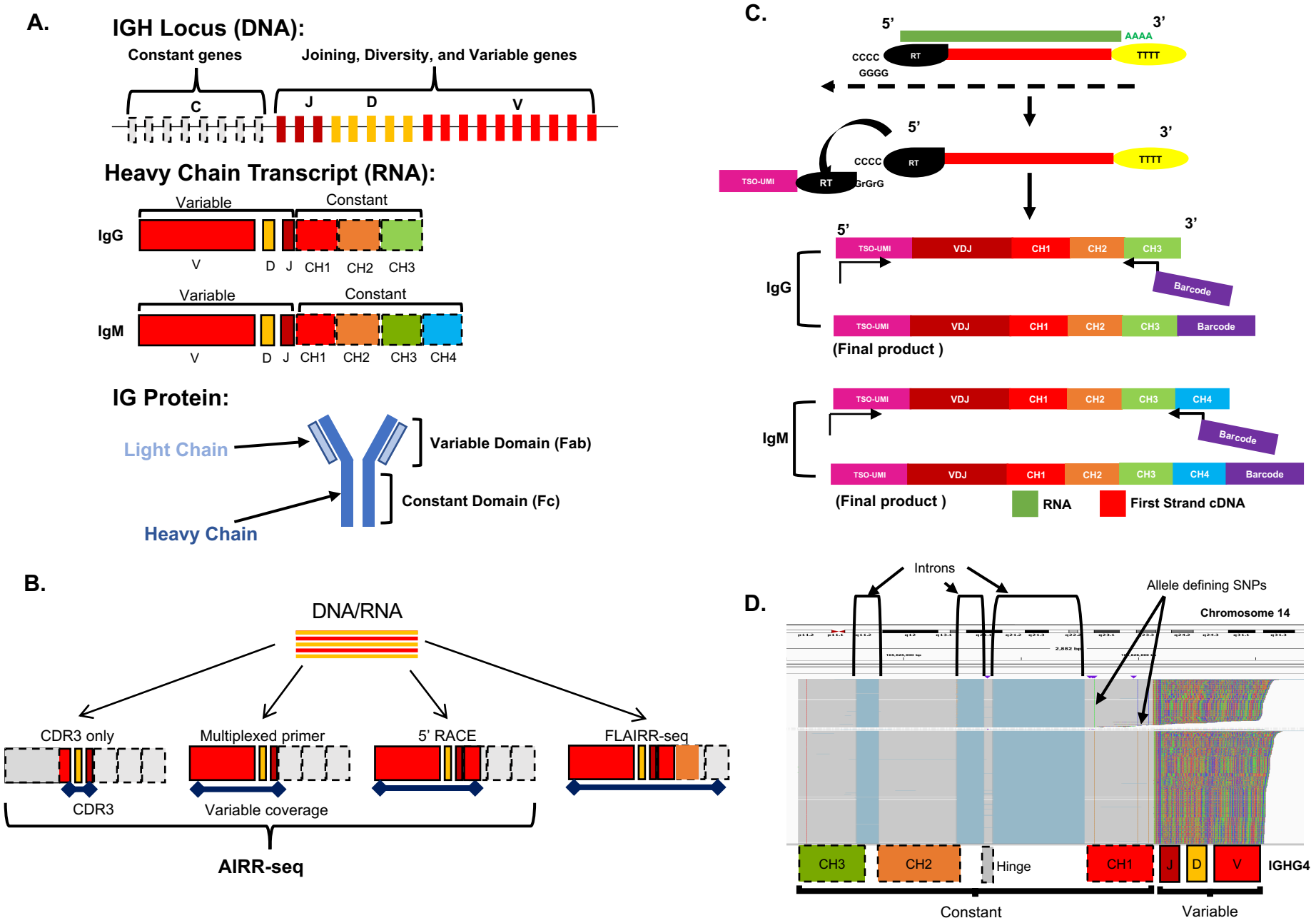

A.
IGH Locus (DNA):
Constant genes
Joining, Diversity, and Variable genes
C
J
D
V
Heavy Chain Transcript (RNA):
Variable
Constant
IgG
V
D
J
CH1
CH2
CH3
Variable
Constant
IgM
V
D
J
CH1
CH2
CH3
CH4
IG Protein:
Light Chain
Heavy Chain
Variable Domain (Fab)
Constant Domain (Fc)
B.
DNA/RNA
CDR3 only
Multiplexed primer
5' RACE
FLAIRR-seq
CDR3
Variable coverage
AIRR-seq
C.
5'
3'
AAAA
CCCC
GGGG
RT
TTTT
TSO-UMI
GrGrG
VDJ
CH1
CH2
CH3
CH4
Barcode
IgG
IgM
(Final product)
RNA
First Strand cDNA
D.
Introns
Allele defining SNPs
Chromosome 14
CH3
CH2
Hinge
CH1
J
D
V
IGHG4
Constant
Variable

FIGURE 2.

A. IgG BC vs. PBMC FLAIRR-seq

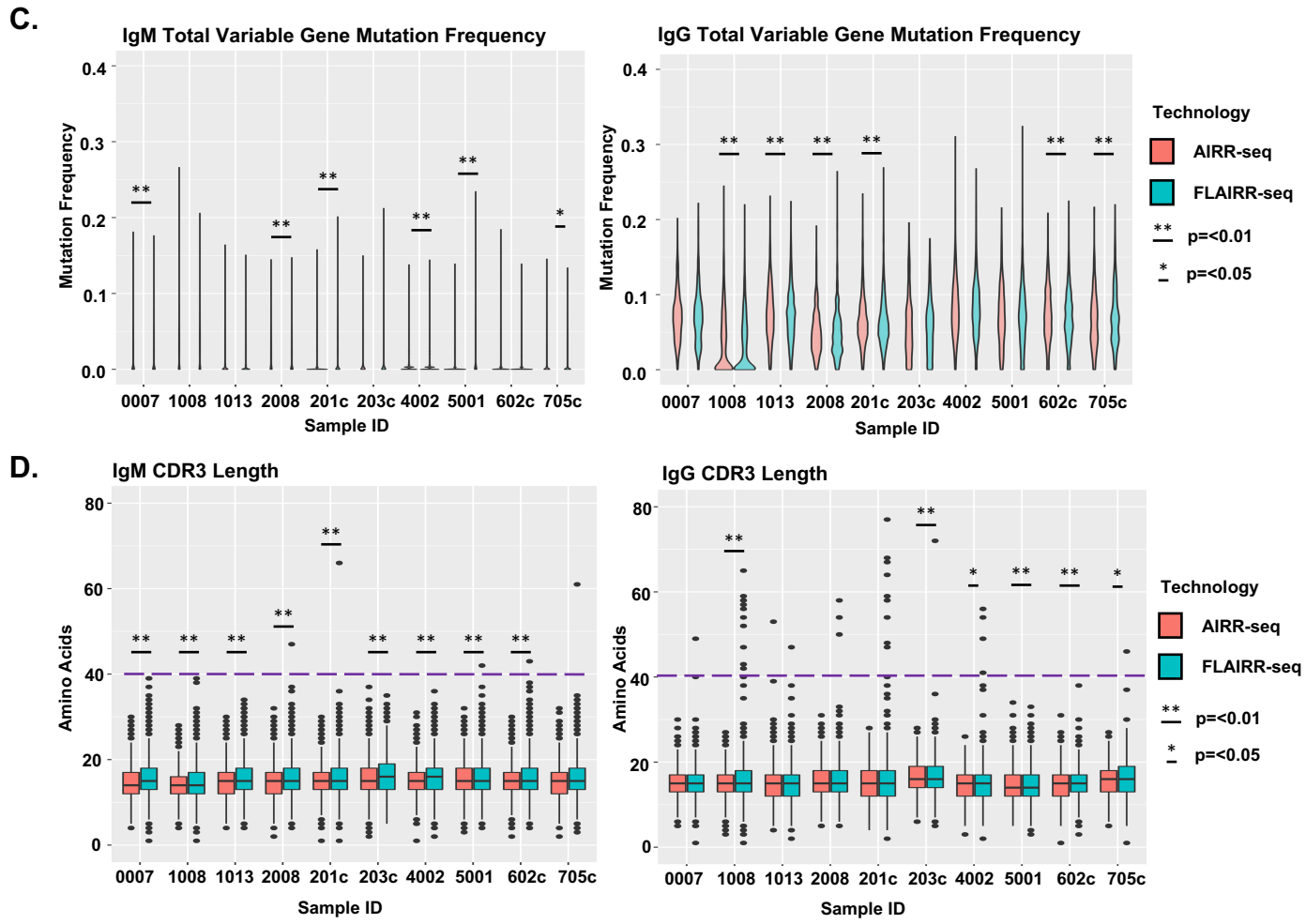
	V		D		J	
	r	p	r	p	r	p
1008	0.95	2.2e−16	0.95	2.2e−16	1.00	0.0028
0007	0.95	2.2e−16	0.95	4.1e−16	0.94	0.017
4002	0.95	2.2e−16	0.98	2.2e−16	0.89	0.033
705c	0.90	2.2e−16	0.97	2.2e−16	0.94	0.017

B. 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



E. 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.

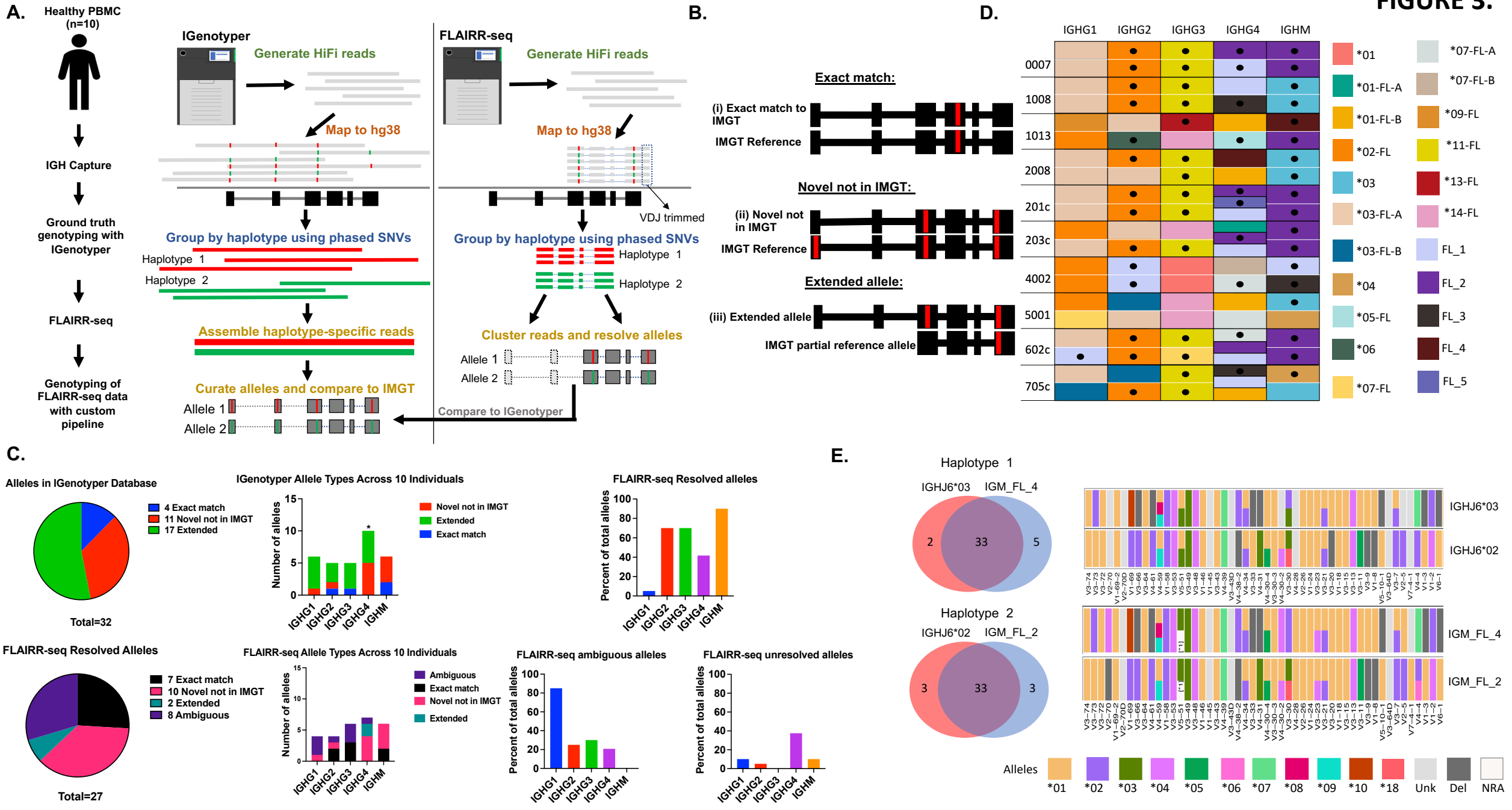

A.
Healthy PBMC (n=10)
IGH Capture
Ground truth genotyping with IGenotyper
FLAIRR-seq
Genotyping of FLAIRR-seq data with custom pipeline
IGenotyper
Generate HiFi reads
Map to hg38
Group by haplotype using phased SNVs
Haplotype 1
Haplotype 2
Assemble haplotype-specific reads
Curate alleles and compare to IMGT
Allele 1
Allele 2
FLAIRR-seq
Generate HiFi reads
Map to hg38
VDJ trimmed
Group by haplotype using phased SNVs
Haplotype 1
Haplotype 2
Cluster reads and resolve alleles
Allele 1
Allele 2
Compare to IGenotyper
B.
Exact match:
(i) Exact match to IMGT
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
IGHM
0007
1008
1013
2008
201c
203c
4002
5001
602c
705c
*01
*01-FL-A
*01-FL-B
*02-FL
*03
*03-FL-A
*03-FL-B
*04
*05-FL
*06
*07-FL
*07-FL-A
*07-FL-B
*09-FL
*11-FL
*13-FL
*14-FL
FL_1
FL_2
FL_3
FL_4
FL_5
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
Novel not in IMGT
Extended
Exact match
IGHG1
IGHG2
IGHG3
IGHG4
IGHM
FLAIRR-seq Resolved alleles
Percent of total alleles
FLAIRR-seq Resolved Alleles
7 Exact match
10 Novel not in IMGT
2 Extended
8 Ambiguous
Total=27
FLAIRR-seq Allele Types Across 10 Individuals
Number of alleles
Ambiguous
Exact match
Novel not in IMGT
Extended
FLAIRR-seq ambiguous alleles
Percent of total alleles
FLAIRR-seq unresolved alleles
Percent of total alleles
E.
Haplotype 1
IGHJ6*03
IGM_FL_4
2
33
5
Haplotype 2
IGHJ6*02
IGM_FL_2
3
33
3
IGHJ6*03
IGHJ6*02
IGM_FL_4
IGM_FL_2
Alleles
*01
*02
*03
*04
*05
*06
*07
*08
*09
*10
*18
Unk
Del
NRA

FIGURE 4.

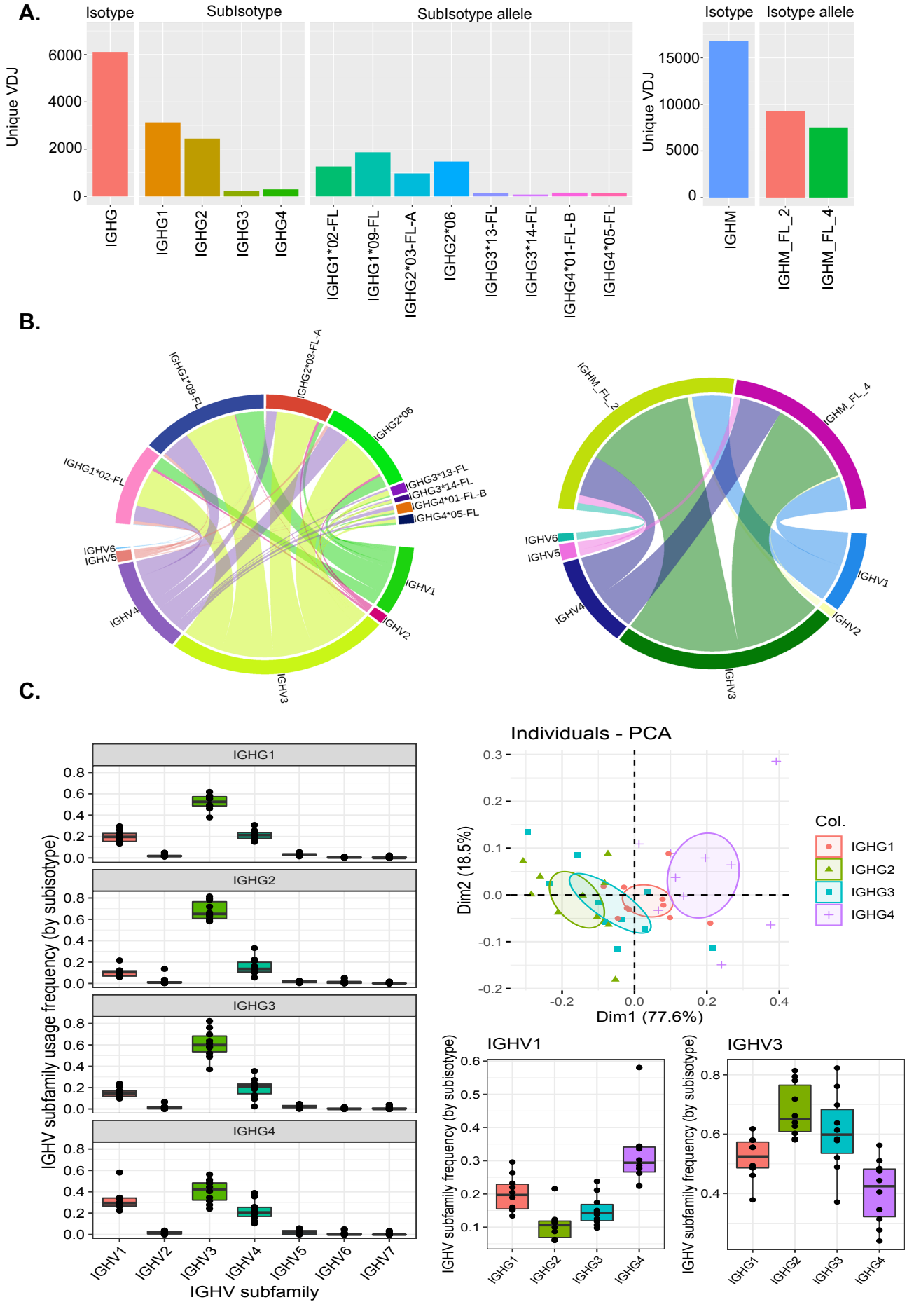
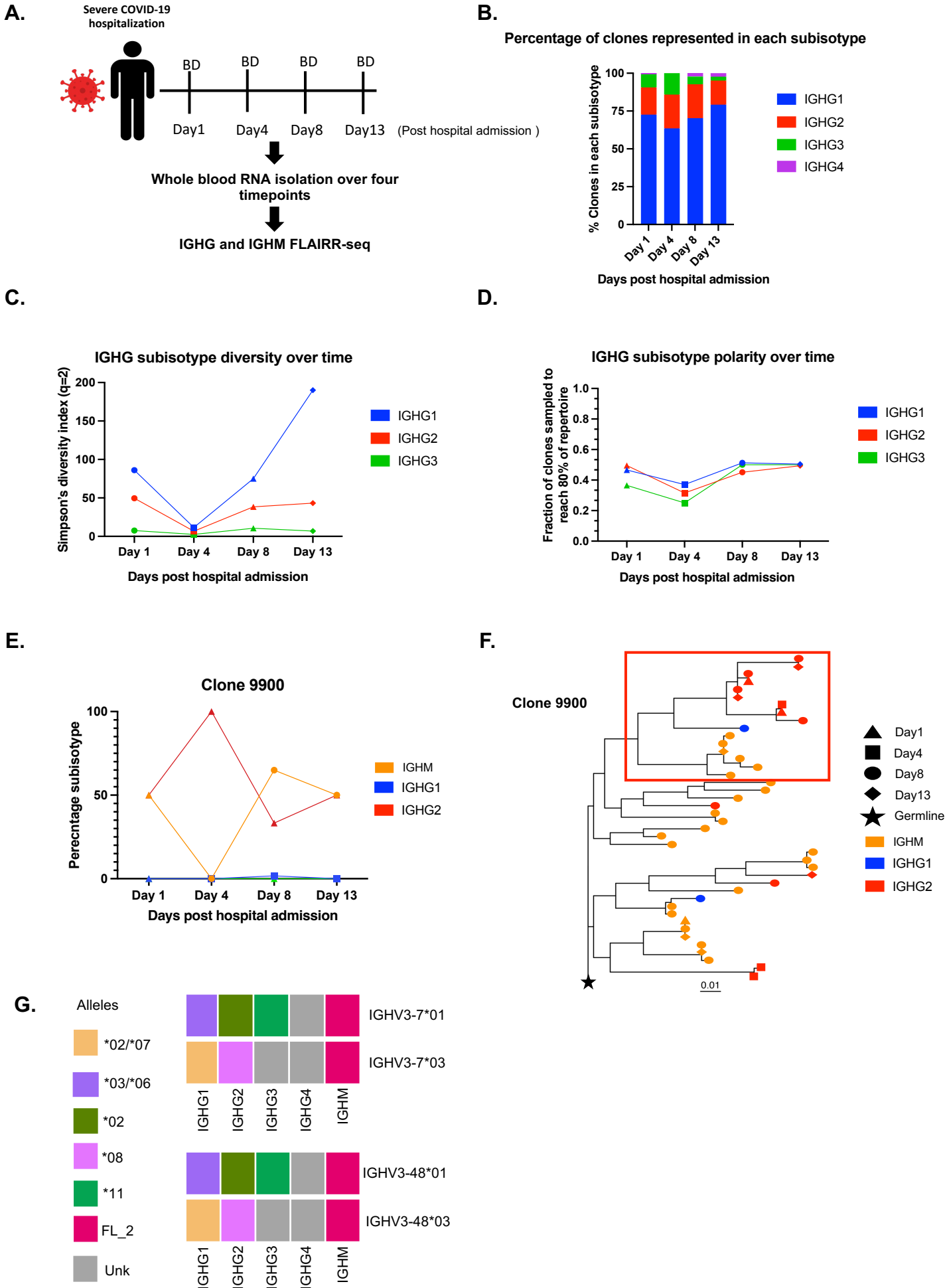

A.
Isotype
SubIsotype
SubIsotype allele
Isotype
Isotype allele
Unique VDJ
IGHG
IGHG1
IGHG2
IGHG3
IGHG4
IGHG1*02-FL
IGHG1*09-FL
IGHG2*03-FL-A
IGHG2*06
IGHG3*13-FL
IGHG3*14-FL
IGHG4*01-FL-B
IGHG4*05-FL
IGHM
IGHM_FL_2
IGHM_FL_4
B.
IGHV1
IGHV2
IGHV3
IGHV4
IGHV5
IGHV6
C.
IGHG1
IGHG2
IGHG3
IGHG4
IGHV subfamily usage frequency (by subisotype)
IGHV1
IGHV2
IGHV3
IGHV4
IGHV5
IGHV6
IGHV7
IGHV subfamily
Individuals - PCA
Dim2 (18.5%)
Dim1 (77.6%)
Col.
IGHV subfamily frequency (by subisotype)

FIGURE 5.


A.
Severe COVID-19 hospitalization
BD
BD
BD
BD
Day1
Day4
Day8
Day13
(Post hospital admission)
Whole blood RNA isolation over four timepoints
IGHG and IGHM FLAIRR-seq
B.
Percentage of clones represented in each subisotype
% Clones in each subisotype
100
75
50
25
0
Day 1
Day 4
Day 8
Day 13
Days post hospital admission
IGHG1
IGHG2
IGHG3
IGHG4
C.
IGHG subisotype diversity over time
Simpson's diversity index (q=2)
200
150
100
50
0
Day 1
Day 4
Day 8
Day 13
Days post hospital admission
IGHG1
IGHG2
IGHG3
D.
IGHG subisotype polarity over time
Fraction of clones sampled to reach 80% of repertoire
1.0
0.8
0.6
0.4
0.2
0.0
Day 1
Day 4
Day 8
Day 13
Days post hospital admission
IGHG1
IGHG2
IGHG3
E.
Clone 9900
Perecntage subisotype
100
50
0
Day 1
Day 4
Day 8
Day 13
Days post hospital admission
IGHM
IGHG1
IGHG2
F.
Clone 9900
0.01
Day1
Day4
Day8
Day13
Germline
IGHM
IGHG1
IGHG2
G.
Alleles
*02/*07
*03/*06
*02
*08
*11
FL_2
Unk
IGHV3-7*01
IGHV3-7*03
IGHV3-48*01
IGHV3-48*03
IGHG1
IGHG2
IGHG3
IGHG4
IGHM