# Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing

Charles Cole,[1] Ashley Byrne,[2] Matthew Adams,[2] Roger Volden,[1] and Christopher Vollmers[1]

[1]*Department of Biomolecular Engineering, Cellular,* [2]*Department of Molecular, Cellular, and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA*

**The human immune system relies on highly complex and diverse transcripts and the proteins they encode. These include transcripts encoding human leukocyte antigen (HLA) receptors as well as B cell and T cell receptors (BCR and TCR). Determining which alleles an individual possesses for each HLA gene (high-resolution HLA typing) is essential to establish donor–recipient compatibility in organ and bone marrow transplantations. In turn, the repertoires of millions of unique BCR and TCR transcripts in each individual carry a vast amount of health-relevant information. Both short-read RNA-seq-based HLA typing and BCR/TCR repertoire sequencing (AIRR-seq) currently rely on our incomplete knowledge of the genetic diversity at HLA and BCR/TCR loci. Here, we generated over 10,000,000 full-length cDNA sequences at a median accuracy of 97.9% using our nanopore sequencing-based Rolling Circle Amplification to Concatemeric Consensus (R2C2) protocol. We used this data set to (1) show that deep and accurate full-length cDNA sequencing can be used to provide isoform-level transcriptome analysis for more than 9000 loci, (2) generate accurate sequences of HLA alleles, and (3) extract detailed AIRR data for the analysis of the adaptive immune system. The HLA and AIRR analysis approaches we introduce here are untargeted and therefore do not require prior knowledge of the composition or genetic diversity of HLA and BCR/TCR loci.**

**[Supplemental material is available for this article.]**

The human immune system relies on highly diverse and complex receptors to protect us from a wide array of pathogens. The transcripts encoding these immune receptors are of great interest to basic and translational research as well as diagnostic and other clinical purposes (Logan et al. 2011; Weng et al. 2013; Vollmers et al. 2015). However, RNA-seq, the current gold-standard for whole-transcriptome analysis, falls short of describing these immune receptors completely and accurately (Mose et al. 2016; Bolotin et al. 2017). Accurate and deep full-length cDNA sequencing of immune cell transcriptomes could overcome this shortfall by providing (1) the isoforms of surface receptors targeted in immunotherapy, (2) allele-resolved HLA transcript sequences central to self/non-self-discrimination, and (3) B cell receptor (BCR) and T cell receptor (TCR) repertoires instrumental to the adaptive immune response to pathogens.

First, full-length cDNA sequencing should be capable of investigating the transcript isoforms of surface receptors expressed by B cells which have important roles in the immune response but are also themselves targets in the treatment of B cell–derived leukemia. For example, current antibody and chimeric antigen receptor (CAR) T cell therapies against B cell acute lymphoblastic leukemia (B-ALL) target epitopes of CD19, CD20, and CD22 (Davila and Brentjens 2016; Fry et al. 2018). However, evidence is accumulating that these epitopes might be absent in some isoforms expressed by these genes (Sotillo et al. 2015; Byrne et al. 2017; Fischer et al. 2017). Determining isoform-level transcriptomes of healthy and cancerous immune cells might inform treatment decisions and future development.

Second, full-length cDNA sequencing could be used to accurately determine the sequence and identity of alleles of HLA genes. HLA genes encode for the major histocompatibility complex (MHC) group of immune receptors which are instrumental in the presentation of antigens on the cell surface (Shiina et al. 2009). Determining the identities of HLA alleles that are present within an individual's genome (HLA typing) is of central importance to establish compatibility of donor and recipient for organ and bone marrow transplantation.

Currently, HLA typing is performed in clinical laboratories by amplifying the genomic DNA encoding these genes and sequencing the resulting amplicons with short-read sequencers (Wang et al. 2012). This targeted DNA-based approach is required because even sophisticated computational tools relying on complex workflows and statistically rigorous frameworks struggle to process RNA-seq data and provide reliable allelic identities (Boegel et al. 2012; Orenbuch et al. 2019). Extracting accurate and full-length HLA allele sequences from full-length cDNA sequences could therefore simplify the computational task of determining the identities of HLA alleles present in a sample.

Third, accurate full-length cDNA sequencing should also be capable of determining adaptive immune receptor sequences,

**Corresponding author: vollmers@ucsc.edu**
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.257188.119.

including BCR heavy and light chains and TCR alpha and beta transcripts. These receptor transcripts contain a constant region that determines the type and characteristics of the receptor and a variable region that determines its binding affinity. The exon encoding the variable regions is generated through the process of somatic VDJ recombination, which randomly recombines one gene segment each from pools of similar but distinct V, D, and J gene segments (Tonegawa 1983). Each developing T or B cell uses this somatic recombination process unique to these cell types to rearrange nonfunctional loci into two functional genes (B cells: heavy/light [kappa or lambda], T cells: alpha/beta). The repertoires of these transcripts present in blood or tissue samples carry a large amount of information on the composition of the loci they are expressed from, as well as the activation state, clonal composition (including malignant clones in leukemia) (Logan et al. 2011), and basic biological processes of the adaptive immune system.

Currently, targeted adaptive immune receptor repertoire sequencing assays (AIRR-seq) methods are routinely used to sequence these transcripts and investigate the human immune system (Horns et al. 2016; de Bourcy et al. 2017a,b). Specialized assays are required for this task because the diversity and unique nature of these transcripts make them practically impossible to analyze at full length using standard RNA-seq protocols. Further, the majority of these assays are based on primers against known V segments and therefore are potentially biased against so far unknown V segments. The ability to instead extract these full-length transcripts in an unbiased way from whole-transcriptome full-length cDNA sequencing would greatly simplify workflows and expand the information that can be recovered from nontargeted transcriptome analysis.

With full-length transcriptome sequencing rapidly maturing (Sharon et al. 2013; Oikonomopoulos et al. 2016; Gupta et al. 2018; Workman et al. 2019), we wanted to investigate its current potential for these types of analyses. Here, we use our previously published R2C2 method (Volden et al. 2018; Byrne et al. 2019) implemented on the Oxford Nanopore Technologies MinION sequencer to analyze RNA extracted from a human peripheral blood mononuclear cells (PBMC) sample—a mix of mostly monocytes, B cells, and T cells. We analyzed the resulting data to identify transcript isoforms, allele-resolved sequences of full-length HLA transcripts, as well as extract repertoires of BCR and TCR transcripts.

Our results show that accurate and deep full-length cDNA sequencing can resolve the most complex transcripts in the mammalian genome and represents an unbiased alternative to current HLA typing and AIRR-seq methods.

## Results

We extracted total RNA and genomic DNA from PBMC samples purified from the whole blood of a healthy male adult. DNA was used for high-resolution HLA typing, whereas total RNA was used for several transcriptome analysis assays. First, we generated full-length cDNA using a modified first half of the Smart-seq2 protocol (Picelli et al. 2014b). cDNA was then split to generate sequencing libraries with three different methods. First, to generate standard Smart-seq2 libraries, part of this cDNA was tagmented using Tn5 (Picelli et al. 2014a). Next, we circularized the full-length cDNA and performed rolling circle amplification on the resulting circular DNA. This reaction generated long dsDNA containing multiple concatemeric copies of the original full-length cDNA. We either sequenced this DNA on the ONT MinION using the R2C2 protocol (Volden et al. 2018) or tagmented it with Tn5 to generate a hybrid Smart-seq2 and R2C2 short-read library we named Smar2C2 (Fig. 1). Because of their distinct features, the different libraries were intended for different tasks. R2C2 data were intended to provide full-length cDNA sequences for isoform identification and AIRR characterization. Smart-seq2 and Smar2C2 data were intended to provide highly accurate short-read data for SNP identification and isoform sequence polishing, with Smar2C2 data providing improved coverage of transcript ends.

### R2C2 data characteristics

The R2C2 method sequences the same cDNA sequence multiple times to overcome the low accuracy of raw 1D ONT cDNA reads (Workman et al. 2019). In an update to the previously published version (Volden et al. 2018) of R2C2, we also incorporated the use of unique molecular identifiers (UMIs) as part of the DNA splints. Because the UMIs are linked to cDNA after PCR amplification they do not indicate unique RNA molecules, but instead reflect unique circularization events, thereby allowing us to combine the raw reads originating from the same cDNA molecule to improve R2C2 consensus read accuracy.

We generated R2C2 data in four technical replicates using individual ONT MinION 9.4.1 flow cells. 1D raw reads from each flow cell were processed using the C3POa program (Volden et al. 2018) to generate a total of 10,298,086 R2C2 consensus reads (Fig. 2A). Of these reads, 122,353 could be grouped with one or more other reads based on their UMIs and were combined into 58,893 R2C2-UMI reads. Analysis of the resulting reads showed an accuracy of 97.9% for regular R2C2 reads and of 99.3% for R2C2-UMI reads (Fig. 2B). Because of the relatively low number of these R2C2-UMI reads, we merged regular R2C2 reads and R2C2-UMI reads into a single data set with a total number of 10,234,626 reads with a median length of 721 nt (Fig. 2C) of which >99.9% aligned to the human genome.

We used featureCounts (Liao et al. 2014) to quantify gene-level expression based on these R2C2 reads and Smart-seq2 reads. Gene expression as determined by the different protocols showed a Pearson's r of 0.93 suggesting good correlation between them. Although this suggests that R2C2 captures the transcriptome in a quantitative manner, it is not the focus of this analysis as the strength of long-read sequencing like R2C2 is not gene-level but instead isoform-level transcriptome analysis.

### Isoform identification and evaluation

To identify transcript isoforms, we used the R2C2 reads as input into a revised version of the Mandalorion program (v3, GitHub) (Byrne et al. 2017). We detected 21,358 transcript isoforms expressed from 9971 gene loci with a median length of 1152 nt (Fig. 2C). The isoform sequences Mandalorion produced showed a median accuracy of 99.7%. Their actual accuracy is likely even higher considering the genome sequence of the sample donor is not expected to be identical to the human reference genome sequence.

We further evaluated the quality of these isoforms using SQANTI (Fig. 2A; Tardaguila et al. 2018), by which 12,250 isoforms were scored as "full-splice match" (FSM) of an annotated transcript; 2265 isoforms were scored as "novel in catalog" (NIC), meaning that they used annotated splice sites in previously unannotated combinations; 1369 isoforms were scored as "novel not in catalog" (NNC), which means that they contain unannotated

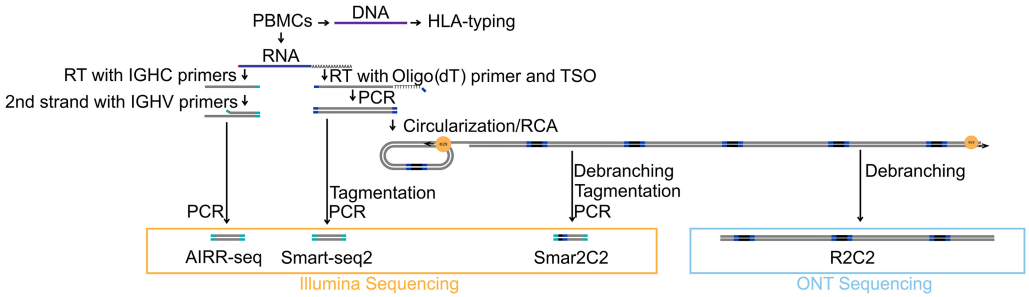


**Figure 1.** Analysis of the adaptive immune system through high-throughput sequencing. Schematic of experiment design. DNA and RNA extracted from a PBMC sample underwent several library preparation protocols to generate AIRR-seq, Smart-seq2, Smar2C2, and R2C2 libraries that were sequenced on Illumina and ONT sequencers.

splice sites and potentially entirely unannotated exons; and 4661 isoforms were scored as "incomplete splice match" (ISM), which could mean that they are potential artifacts.

To validate isoform features (exons, 5′ ends, 3′ ends) that did not match the GENCODE (v29) annotation (Harrow et al. 2012), we used a new short-read Illumina protocol. Because all general purpose RNA-seq protocols struggle to capture the ends of transcripts, we developed Smar2C2, which is a hybrid of the Smart-seq2 and R2C2 methods that tagments not cDNA molecules but the cDNA concatemers generated as part of the R2C2 method (Fig. 1). As a result, Tn5-based tagmentation is not affected by cDNA ends because these ends are now encapsulated within a much larger DNA molecule. Gene expression as determined by Smar2C2 and Smart-seq2 showed a Pearson's r of 0.97 suggesting that the Smar2C2 protocol does not distort the cDNA composition (Supplemental Fig. S1). Further, the Smar2C2 protocol was more likely than the Smart-seq2 protocol to cover transcript ends in the form of reads containing oligo sequences used in cDNA generation. Reads containing template switch oligo sequences (TSO reads) cover transcript 5′ ends and reads containing Oligo(dT) sequences (Oligo[dT] reads) cover transcript ends 3′ ends (Methods). Processing of the approximately 39 million Smart-seq2 and 23 million Smar2C2 read pairs suggests that Smar2C2 data contain about 6× (Smart-seq2: 3% and Smar2C2: 17% of all reads) more TSO reads and about 25× (Smart-seq2: 0.17% and Smar2C2: 4% of all reads) more Oligo(dT) reads than Smart-seq2 data.

Smar2C2 read coverage dropped sharply outside the splice sites of 365 exons not overlapping any exons in the GENCODE annotation. Further, TSO and Oligo(dT) read coverage dropped sharply outside of 5525 5′ ends and 5712 3′ ends of FSM, NIC, and NNC isoforms that were more than 10 nt away from annotated transcription start sites (TSS) and poly(A) sites, respectively. An *x*-nt distance to a GENCODE TSS does not guarantee a biologically distinct element. The distance cutoff of 10 nt and the number of features identified with this cutoff are therefore ultimately arbitrary and do not carry biological significance. Together, this indicated that isoform features identified by Mandalorion were indeed present within our cDNA pool even if they did not match the GENCODE annotation (Fig. 2D).

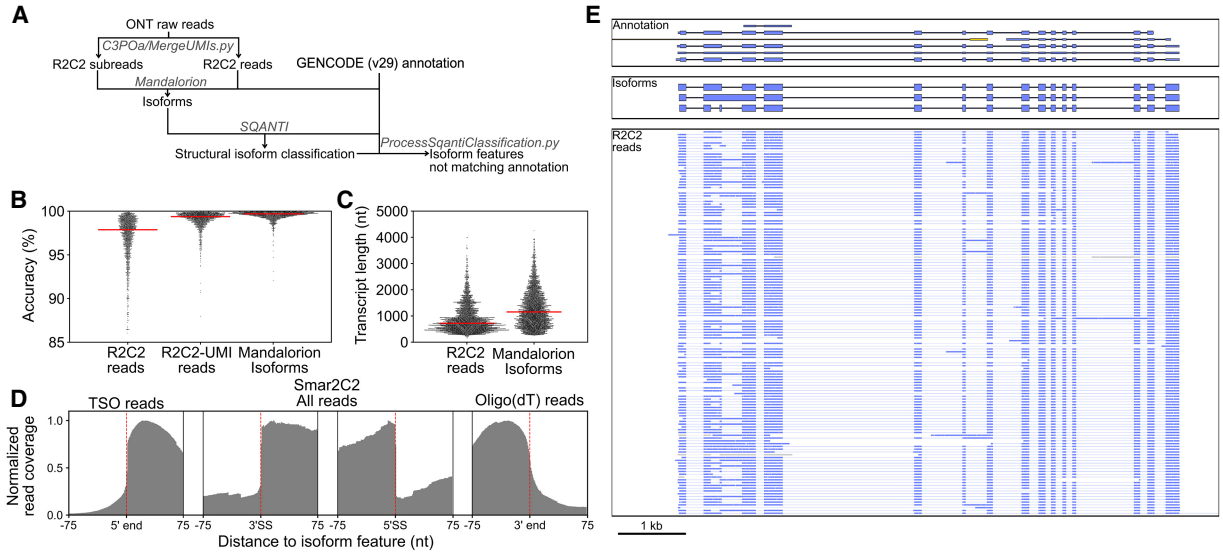


**Figure 2.** Isoform-level transcriptome analysis. (*A*) Flow chart of the isoform analysis pipeline. (*B*,*C*) Swarm plots of the (*B*) accuracy [*Matches/(Matches + Mismatches + Indels)*] and (*C*) transcript length [*Matches + Mismatches*] of R2C2 reads (with and without UMI) and isoforms. Red line indicates median. (*D*) Normalized Smar2C2 read coverage around isoform features (exons, 5′/3′ ends) that did not match GENCODE annotation determined by either all (splice sites), TSO (5′ ends), or Oligo(dT) (3′ ends) reads. (*E*) The *CD19* gene locus is shown in a genome browser view. GENCODE annotation (*top*), Mandalorion isoforms (*center*), and R2C2 reads (*bottom*) are shown. Direction of shown features is indicated by color ("top strand": blue, "bottom strand": yellow).

To showcase the usefulness of deep full-length isoform data we highlight the *CD19* gene expressed by B cells. *CD19* is of great importance because the protein it encodes is a target of cancer immunotherapy in B cell acute lymphoblastic leukemia (ALL) and other B cell cancers. At a total read depth of approximately 10 million reads, only 146 reads aligned to the *CD19* gene. These 146 R2C2 reads appeared to split into three major and several minor isoforms (Fig. 2E). Mandalorion identified the three major isoforms that confirmed a previously identified splice site in the second exon of *CD19* that we recently observed in single B cells (Volden et al. 2018). Because they encode distinct proteins, these isoforms may affect whether CAR T cells (Sotillo et al. 2015; Fischer et al. 2017) can bind leukemia cells expressing them.

## Allele-specific isoform expression

At close to 98% accuracy, R2C2 reads should be well suited for allele-specific isoform expression analysis. Because the SNP identification is much more established with short-read data, we made use of the Smart-seq2 and Smar2C2 data that we produced to identify SNPs present in the PBMC sample we analyzed using the standard genome analysis toolkit (GATK) RNA-seq workflow (Fig. 3A; Van der Auwera et al. 2013). Using the TurboPhaser.py (Methods) script, we then extracted heterozygous SNPs from this list and phased these SNPs within gene boundaries using R2C2 reads. TurboPhaser.py then sorted R2C2 reads and short-read RNA-seq read pairs into alleles based on the phased SNPs they contained. Overall, we assigned 756,072 R2C2 reads (7.4% of all reads, Allele1: 377,794; Allele2: 378,278) and 1,817,151 RNA-seq read pairs (2.9% of all read pairs, Allele1: 872,130; Allele2: 945,021) to either of two alleles. It is noteworthy that R2C2 reads are more than twice as likely to be sorted into alleles than RNA-seq read pairs based on the same set of phased SNPs. This is likely caused by R2C2 reads, in contrast to RNA-seq reads, covering entire transcripts and all SNPs they contain.

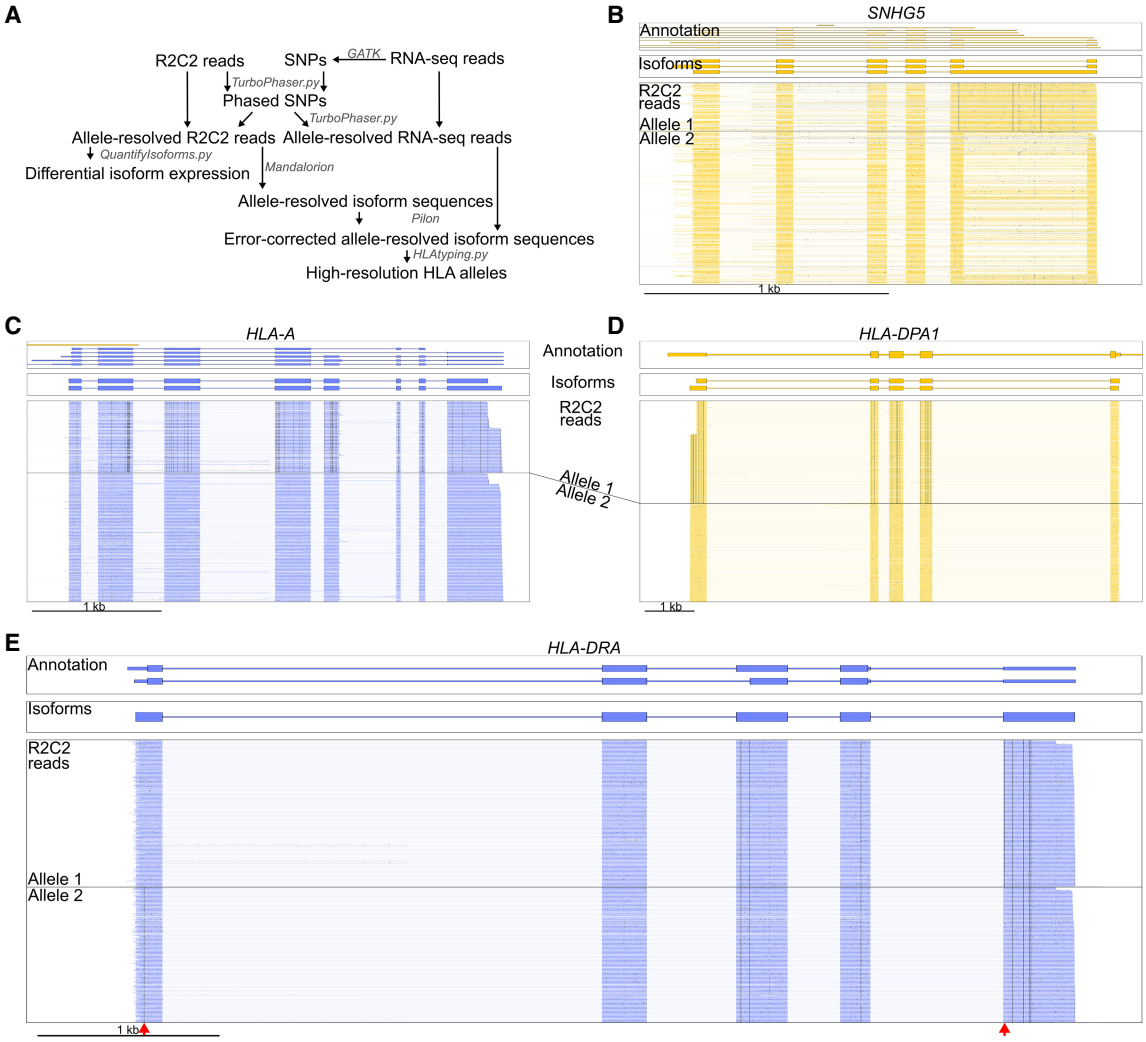


**Figure 3.** Allele-resolved isoform expression and sequences. (*A*) Computational strategy for determining allele-resolved isoforms. (*B–E*) *SNHG5*, *HLA-A*, *HLA-DPA1*, and *HLA-DRA* gene loci are shown in a genome browser view. GENCODE annotation (*top*), Mandalorion isoforms (*center*), and allele-resolved R2C2 reads (*bottom*) are shown. Direction of shown features is indicated by color ("top strand": blue, "bottom strand": yellow). Mismatches of R2C2 reads to the genome reference are shown in black. Both *HLA-A* and *HLA-DPA* show allele-resolved differential expression of isoforms with alternative poly(A) sites. (*E*) Red arrows indicate allele-specific variants in the *HLA-DRA* gene.

Next, we used the allele-resolved R2C2 reads to quantify the expression of the previously identified isoforms. To this end, we separated these reads into the four technical replicates based on the flow cells they were generated on. Using DESeq (Anders and Huber 2012), we then identified roughly 80 isoforms that showed differential expression between the two alleles while accounting for the technical variation associated with each minion run (Supplemental Table S2). The *SNHG5* gene highlights this differential expression with transcripts of one allele always retaining the first intron of the gene and transcripts of the other allele either splicing or retaining the intron (Fig. 3B). Seven of the roughly 80 differentially expressed isoforms originated from an HLA gene. *HLA-A* and *HLA-DPA1* both show differentially expressed isoforms with alternative poly(A) sites. Although the alternative *HLA-A* isoform was previously observed (Kulkarni et al. 2017a,b), the *HLA-DPA1* isoform was not (Fig. 3C,D).

## Allele-resolved isoform sequences enable high-resolution HLA typing

Next, we investigated whether allele-resolved R2C2 reads are suited for identifying which HLA alleles are present in an individual. Current RNA-seq-based HLA typing methods rely on databases of previously identified and systematically cataloged HLA alleles. The IPT-IMGT/HLA database contains the systematic names and sequences of thousands of different HLA alleles. The systematic names (e.g., HLA-A*01:01:01:01) contain multiple groups of digits separated by colons to denote the relationship between sequences.

Using this database, current RNA-seq based methods can determine the identity of HLA alleles present in an individual with 4- to 6-digit resolution. However, even the most advanced methods like arcasHLA (Orenbuch et al. 2019) have a 10% error-rate for the identification of some HLA genes and cannot determine new HLA alleles absent from the database they use. Reliable HLA typing therefore still requires dedicated DNA-based approaches. These approaches PCR-amplify full-length HLA genes from genomic DNA and determine the sequence of the resulting amplicons. These sequences are then compared to the database of known HLA alleles to determine 6-digit HLA types.

To test whether R2C2 would enable a similar approach, we used the 377,794 R2C2 reads assigned to the first allele and 378,278 R2C2 reads assigned to the second allele as separate inputs into Mandalorion, which then generated 2237 and 2056 allele-specific isoforms, respectively. Mandalorion generates entirely read-based consensus sequences for each isoform it identifies, which in this case included at least one full-length isoform for each major HLA gene on either allele. Next, to achieve the highest possible accuracy required for identifying variants and achieving unambiguous HLA typing, we used allele-resolved RNA-seq reads to error-correct the allele-specific Mandalorion isoforms with Pilon (Walker et al. 2014).

To determine the identity of the HLA alleles present in the analyzed sample, we used the HLAtyping.py utility of Mandalorion that aligns these error-corrected allele-resolved HLA isoform sequences to the complete database of HLA alleles (Robinson et al. 2000, 2015) using minimap2 and extracts the best match for each HLA gene. After finding the best HLA allele match of an error-corrected allele-specific isoform, we truncated the match to 6-digit resolution to match clinical high-resolution HLA typing. All HLA alleles we identified in this way (R2C2 + RNA-seq) matched DNA-based high-resolution HLA typing performed at the Immunogenetics and Transplantation Laboratory (ITL) at UCSF (Targeted Amplicon NGS) (Table 1).

Having confirmed the accuracy of the HLA alleles we identified, we compared them to HLA alleles determined using only RNA-seq data and different programs. The seq2HLA package (RNA-seq/seq2HLA) only generates data with 4-digit resolution and failed to identify *HLA-DPA1*, *HLA-DQA1*, and *HLA-DQB1* as heterozygous and miscalled *HLA-DPB1*. The arcasHLA (Orenbuch et al. 2019) package (RNA-seq/arcasHLA) performed better, determining the correct HLA alleles for all HLA genes. However, although both RNA-seq/arcasHLA and R2C2 + RNA-seq strategies identify only one 6-digit resolution allele for the *HLA-DRA* genes (HLA-DRA*01:01:01), only R2C2 + RNA-seq identifies the *HLA-DRA* gene as heterozygous by identifying distinct sequences for the two alleles. Because targeted amplicon NGS data were not available for *HLA-DRA* to serve as ground truth, we cannot be certain that the R2C2 + RNA-seq strategy yields a correct result for this gene. However, visualizing allele-resolved R2C2 reads suggests the existence of heterozygous variants in two distinct *HLA-DRA* alleles that are several hundred base pairs apart. ArcasHLA cannot resolve this because the variants are outside the protein-coding region, where they affect 8-digit but not 6-digit HLA resolution (Fig. 3E).

Overall, these findings show that accurate full-length cDNA sequencing at high depth allows the determination of highly accurate sequences of HLA alleles, which can then be used for high-resolution HLA typing. In contrast to short-read RNA-seq-based HLA typing, which requires a reference database, these HLA allele sequences can be also used to discover so far unknown HLA alleles.

## Extracting adaptive immune receptor repertoires (AIRR) from R2C2 data

Next, we evaluated whether accurate full-length sequencing could also—completely or in part—replace specialized assays for the analysis of adaptive immune receptor repertoires. To do so, we extracted R2C2 reads from our data set which aligned to BCR or TCR gene loci. Each R2C2 read was treated independently and annotated using the IgBLAST (Ye et al. 2013) algorithm that identifies V, D, and J segments, CDR3 sequences at the V(D)J intersection, and mutations present in each sequence. Finally, we use sequence similarity to determine which constant region is present in each sequence. In this way, we identified tens of thousands of adaptive immune receptor sequences (Table 2).

We then performed in-depth analysis on these annotations with a focus on the transcript sequences encoded by the BCR heavy chain (*IGH*), because the *IGH* locus is the only adaptive immune receptor locus undergoing VDJ recombination, somatic hypermutation, and class-switch recombination (Tonegawa 1983).

We compared R2C2-based *IGH* sequences to 266,390 *IGH* sequences we generated from the same RNA sample by the gold-standard UMI-based targeted AIRR-seq method we developed (called AIRR-seq from here on out) (Vollmers et al. 2013, 2015; Horns et al. 2016, 2019; de Bourcy et al. 2017a). We also used 3261 *IGH* sequences generated from a different RNA sample of the same individual that we published previously (Cole et al. 2016). These sequences were generated using our TMI-seq method which, in contrast to standard AIRR-seq (Vollmers et al. 2013), succeeds in covering the entire V segment by overcoming Illumina read length limitations through a combination of Tn5-based tagmentation and unique molecular identifiers. We focused our analysis on the most relevant features of the *IGH* repertoire, namely (1) CDR3

**Table 1.** R2C2 full-length cDNA sequencing enables high-resolution HLA typing

| | Allele 1 | | | | Allele 2 | | | |
|---|---|---|---|---|---|---|---|---|
| Data source | Targeted amplicon NGS | R2C2 + RNA-seq | RNA-seq | RNA-seq | Targeted amplicon NGS | R2C2 + RNA-seq | RNA-seq | RNA-seq |
| Program | HLA Twin | HLAtyping.py | seq2HLA | arcasHLA | HLA Twin | HLAtyping.py | seq2HLA | arcasHLA |
| *HLA-A* | 03:01:01 | 03:01:01 | 03:01 | 03:01:01 | 32:01:01 | 32:01:01 | 32:01 | 32:01:01 |
| *HLA-B* | 35:01:01 | 35:01:01 | 35:01 | 35:01:01 | 39:01:01 | 39:01:01 | 39:01 | 39:01:01 |
| *HLA-C* | 04:01:01 | 04:01:01 | 04:01 | 04:01:01 | 12:03:01 | 12:03:01 | 12:03 | 12:03:01 |
| *HLA-DRA* | | 01:01:01 | 01:01 | 01:01:01 | | 01:01:01 | 01:01 | |
| *HLA-DRB1* | 16:01:01 | 16:01:01 | 16:01 | 16:01:01 | 01:01:01 | 01:01:01 | 01:01 | 01:01:01 |
| *HLA-DPA1* | 01:03:01 | 01:03:01 | *02:01* | 01:03:01 | 02:01:01 | 02:01:01 | 02:01 | 02:01:01 |
| *HLA-DPB1* | 04:02:01 | 04:02:01 | *105:01* | 04:02:01 | 10:01:01 | 10:01:01 | 10:01 | 10:01:01 |
| *HLA-DQA1* | 01:01:01 | 01:01:01 | *01:02* | 01:01:01 | 01:02:02 | 01:02:02 | 01:02 | 01:02:02 |
| *HLA-DQB1* | 05:01:01 | 05:01:01 | 05:01 | 05:01:01 | 05:02:01 | 05:02:01 | *05:01* | 05:02:01 |

HLA alleles were typed using the programs indicated on *top*. Different programs used for HLA typing rely on different data sources. The HLA Twin program requires DNA-based amplicon sequencing (targeted amplicon NGS), and our HLA-typing.py program requires isoform sequences generated using full-length cDNA (R2C2) and polished with RNA-seq sequences. The seq2HLA and arcasHLA program require only RNA-seq sequences. DRA was not evaluated by targeted amplicon NGS. Contradicting results are shown in italics.

length, (2) isotype usage, (3) V segment composition and usage, (4) somatic hypermutation, and (5) clonality.

### *CDR3 length*

First, we investigated whether R2C2-based *IGH* repertoires capture the full width of CDR3 lengths. The sequences of CDR3s are responsible for the majority of an BCR/antibodies specificity and are composed of semirandom sequence at the intersection of V, D, and J segments. However, CDR3 sequences in functional antibodies are limited to certain lengths to maintain the reading frame between variable and constant regions. This limitation can be clearly observed in CDR3 lengths of AIRR-seq sequences but is less pronounced for R2C2 sequences. This is likely attributable to remaining indel errors in the CDR3 sequences (Fig. 4A). So, although R2C2-based repertoires capture CDR3 sequences of the appropriate length and distribution, downstream analyses that rely on CDR3 length to differentiate functional and nonfunctional *IGH* sequences would be hampered.

### *Isotype usage*

Second, we investigated whether R2C2-based *IGH* repertoires can be used to determine B cell isotype usage. Isotype usage reflects the activity of the adaptive immune system at a given time (Vollmers et al. 2013, 2015; Bashford-Rogers et al. 2019). *IGHM* and *IGHD* sequences are mostly expressed by naive B cells, whereas *IGHA(1-2)*, *IGHG(1-4)*, and *IGHE* sequences are only expressed by previously activated B cells that have undergone class-switch recombination.

Isotype usage was similar between AIRR-seq and R2C2-derived repertoires suggesting that R2C2-derived repertoires will be able to determine the immune activation state of an individual faithfully (Fig. 4B). Improving on any AIRR sequencing approach currently available, R2C2-derived repertoires also resolve whether an *IGH* transcript encodes for a membrane-bound or secreted (antibody) protein. This is possible because R2C2 reads cover the entire 3′ end of *IGH* transcripts where this alternative splicing event occurs. Isoform-level information showed that the majority of *IGHM* transcript are membrane-bound (membrane [M]: 1261; secreted [S]: 417), whereas *IGHD* transcripts were split evenly between the two isoforms (M: 123; S: 164). As expected, over 95% of the sequences of *IGHA(1-2)* and *IGHG(1-4)* isotype subtypes were secreted (e.g., *IGHA1*: M: 182; S: 8113).

### *V segment composition and usage*

Third, we determined whether R2C2 repertoires could be used to investigate V segment usage. Each *IGH* locus is thought to contain 40–50 V segments, and individuals diverge in which V segments and V segment alleles they possess. Standard AIRR-seq assays most often use PCR primers within V segments, thereby masking variation underneath the priming site and missing variation beyond it.

Because the R2C2-based repertoires could not be successfully analyzed with computational tools meant for virtually error-free AIRR-seq (Vander Heiden et al. 2014; Gupta et al. 2015), we determined V segment composition in the analyzed individual by simply counting how many reads were scored by IgBLAST as using a specific V segment with three or fewer mismatches. Reads that assigned equally well to different alleles of the same V segment were counted as ambiguous, whereas reads aligned equally well to different V segments were discarded for this analysis. A V segment allele was counted as detected in a repertoire if it was seen in at least two sequences and accounted for at least 20% of sequences of the V segment

In general, AIRR-seq and R2C2 showed similar recombination frequencies for V segments. However, we found that the deeper AIRR-seq data have an advantage when detecting V segment alleles for V segments that are rarely recombined, including *IGHV1-45*

**Table 2.** Full-length adaptive immune receptor repertoires can be extracted from R2C2 whole transcriptome data

| Receptor | Gene | Number of R2C2 reads |
|---|---|---|
| BCR | *IGH* | 12,863 |
| | *IGL* | 26,759 |
| | *IGK* | 24,460 |
| TCR | *TRA* | 7289 |
| | *TRB* | 13,118 |
| | *TRD* | 316 |
| | *TRG* | 551 |

Number of reads that aligned to the respective locus and could be annotated as AIRR transcripts using IgBLAST is shown.

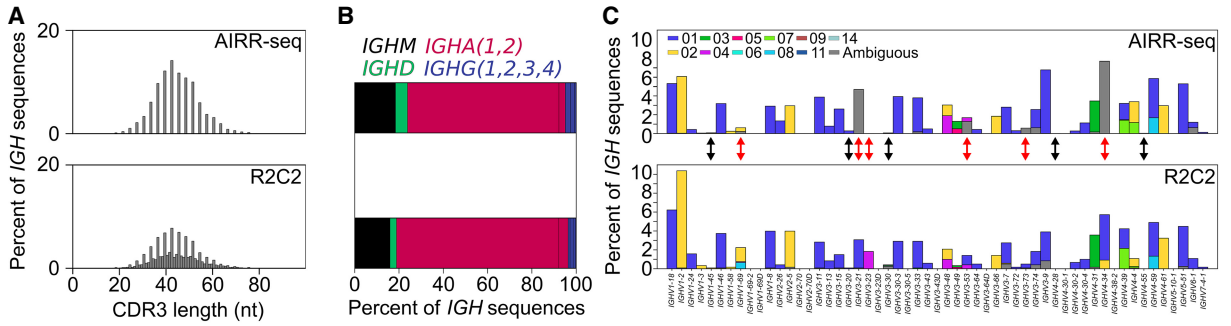


**Figure 4.** R2C2 repertoires have advantages and disadvantages compared to AIRR-seq repertoires: (*A*) CDR3 lengths, (*B*) isotype distribution, and (C) V segment usage for *IGH* repertoires are determined by AIRR-seq (*top*) and R2C2 (*bottom*). In *C*, different bar colors indicate different V segment alleles with gray indicating ambiguous allele calls. Red arrows indicate V segments where AIRR-seq fails to identify an allele unambiguously but R2C2 succeeds. Black arrows indicate rarely recombined V segments that AIRR-seq but not R2C2 detected.

(0.08% of IGH sequences in AIRR-seq data), *IGHV3-20* (0.3%), *IGHV3-30* (0.04%), *IGHV4-28* (0.01%), and *IGHV4-55* (0.02%), which R2C2 did not detect at our requisite abundance of three independent reads. However, we found that R2C2 can unambiguously detect V segments that AIRR-seq could not (Fig 4C). AIRR-seq could not detect one or two of the alleles for *IGHV1-69*, *IGHV3-21*, *IGHV3-23*, *IGHV3-53*, and *IGHV3-73*, whereas R2C2 could. All alleles detected by R2C2 were also detected by TMI-seq data.

This analysis shows that, although specialized AIRR-seq protocols have an edge when detecting V segments that are rarely recombined, it produces incomplete and therefore often ambiguous V segment sequences. Although the TMI-seq method we developed can produce full-length V segment data, it requires a complex workflow and is therefore unlikely to be used for routine clinical analysis. Overall, R2C2 presents an appealing set of trade-offs and is therefore a promising tool for determining the V segment composition and usage within a sample.

### *Somatic hypermutation*

Fourth, we determined whether R2C2 reads would be accurate enough to detect the mutations in *IGH* sequences introduced by somatic hypermutation. We did this by comparing mutations in *IGH* transcript sequences that can undergo somatic hypermutation with *TRB* transcript sequences that are expected to be entirely free of somatic mutations.

We focused this analysis on mismatches that are by far the most common result of somatic hypermutation. We found that R2C2 reads did show only about two mismatches per 300 nt of *TRB* V segment sequence, which corresponds to a mismatch rate of 0.6% and is in line with the remaining 2% total error rate in R2C2 reads being mostly composed of indels. Two mismatches per V segment can therefore be seen as background error in the potential mutated *IGH* sequences we analyzed next. Here, we took advantage of the ability of R2C2 reads to distinguish membrane-bound and secreted (antibodies) isoforms of *IGH* transcripts. *IGHM* transcripts are thought to be mostly expressed by naive unmutated B cells but can also undergo somatic hypermutation. Secreted *IGHM* sequences contained more mutations than membrane-bound *IGHM* sequences, indicating that they are more likely to be expressed by B cells that have undergone activation and somatic hypermutation (Fig. 5A). This difference in mismatch levels disappears in *IGHA1* sequences, which are known to be expressed only by B cells that have undergone activation and somatic hypermutation.

Mismatch levels were significantly higher in R2C2-derived *IGHA1* sequences than in AIRR-seq-derived *IGHA1* sequences (average 24.89 to 20.19, Monte Carlo permutation test $P$-value < 0.00001). Adding randomly sampled R2C2-specific background-level mismatches observed in TRB sequences as well as mismatches observed in the first 20 bases of R2C2 *IGHA1* sequences to AIRR-seq sequences does not abolish this significant difference (average 24.89 to 23.8, Monte Carlo permutation test $P$-value < 0.00001). One possible explanation of this may be that the primers used by AIRR-seq fail to bind highly mutated V segments, which then are not amplified and detected. In turn, this would indicate that R2C2 might have an advantage when investigating highly mutated sequences like those involved in the immune response to HIV (Scheid et al. 2011).

Finally, like AIRR-seq and TMI-seq, *IGHA1* transcript sequenced by R2C2 show the mutational pattern characteristic of somatic hypermutation with mutational hotspots in CDR1 and CDR2. In contrast to AIRR-seq, which uses amplicons primed from FR1 in the *IGH* transcript, TMI-seq and R2C2 sequence detect mutations all the way to the beginning of the V segment (Fig. 5B).

### *Clonality*

Fifth, we investigated the ability of R2C2-based repertoires to capture the clonal composition of B cells in a sample. The *IGH* sequences in a sample can be organized into clones (or lineages), that is, sequences that are expressed by B cells belonging to the same B cell clone. B cell clones originate from a single naive B cell that is activated and starts proliferating. This proliferation is most often associated with somatic hypermutation and class switching. Big lineages are therefore likely to be composed of class-switched sequences with similar but not identical mutation patterns. Also, they have highly similar CDR3 sequences that can be used to group *IGH* sequences into lineages computationally. We performed this analysis for AIRR-seq and R2C2-based repertoires.

The lineages in the two repertoires were closely related (Fig. 6A). One hundred seventy-eight of the top 200 lineages in the AIRR-seq repertoire were also present in the R2C2-based repertoire with missing lineages likely being explained by the lower depth of the R2C2-based repertoire. As expected, most of these repertoires had class-switched to the *IGHA1* isotype and many contained additional lineage-specific mutations.

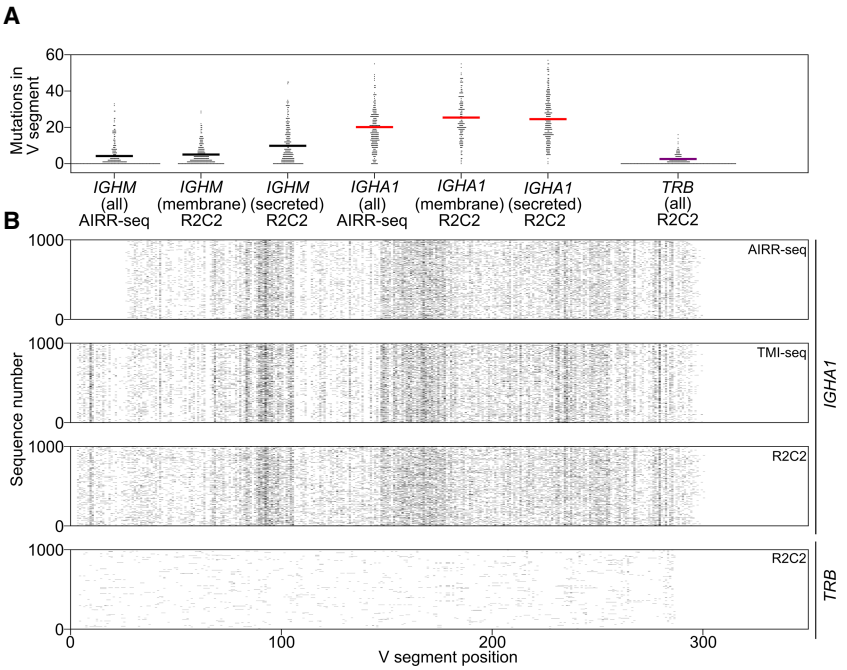


**Figure 5.** Somatic hypermutation can be characterized within R2C2-based IGH repertoires. (*A*) Mismatch mutations per IGH sequence as determined by IgBLAST are shown as swarm plots separated by isotype (*IGHM*, *IGHA1*), isoform (membrane-bound, secreted), and technology (R2C2, AIRR-seq) and compared to *TRB* sequences. Averages are indicated by colored lines. (*B*) The pattern of mutation locations in V segments in AIRR-seq, TMI-seq, and R2C2 sequences is shown for 1000 randomly sampled *IGHA1* or *TRB* sequences.

## Discussion

As genomic analysis of samples becomes an integral component in clinical care, minimizing the number of separate assays that have to be performed and maximizing the information extracted from those assays that are performed should be a top priority. Here, we showcase the potential of our R2C2 full-length cDNA sequencing approach for the in-depth analysis of PMBCs isolated routinely from blood. Smart-seq2 and R2C2 libraries can be generated from <1 ng of total RNA making it possible to use less sample for transcriptome analysis. Further we recently developed a method for the depletion of hemoglobin transcripts from whole blood transcriptomes that would allow this type of analysis to be economically performed on whole blood RNA collected using, for example, PAXgene tubes (Byrne et al. 2019). If implemented on the ONT MinION, as we have done here, R2C2 generates reads of full-length cDNA at 98% accuracy at a cost of about $200 per 1 million reads. We then show that these full-length R2C2 reads can be analyzed with the Mandalorion pipeline to generate transcriptomes that can contain clinically relevant isoform information. R2C2 reads are also suitable to be analyzed with other workflows for isoform determination like FLAIR (Tang et al. 2020), which, using standard settings, requires less read coverage to call isoforms and identifies more single exon isoforms than Mandalorion (Supplemental Table S3).

Beyond isoform annotation and analysis, we show that whole-transcriptome analysis by R2C2 can be used to replace or

Next, we used these mutations to confirm that *IGH* sequences were not spuriously grouped into lineages. To this end, we determined the percentage of mutations in each *IGH* sequence that are shared with other sequences within the same lineage or sequences in different sequences. For both R2C2 and AIRR-seq sequences, this percentage was much higher when comparing sequences within the same lineage confirming the overall accuracy of our lineage grouping approach (Fig. 6B). Finally, to see whether AIRR-seq and R2C2-based repertoires behave similarly when grouped into lineages, we repeatedly subsampled the AIRR-seq repertoire to the depth of the R2C2-based repertoire before grouping the subsampled sequences into lineages (Fig. 6C). The resulting AIRR-seq lineages were slightly larger than R2C2-based lineages at the same depth, indicating that the 2% residual sequencing error present in R2C2 reads causes some related sequences to not be grouped into lineages.

Overall, analysis of AIRR-seq and R2C2-based lineages shows high concordance between the two methods. Of clinical relevance, R2C2-based lineages should be more than capable of tracking B cell clones within and between samples to, for example, track minimal residual disease in leukemia.

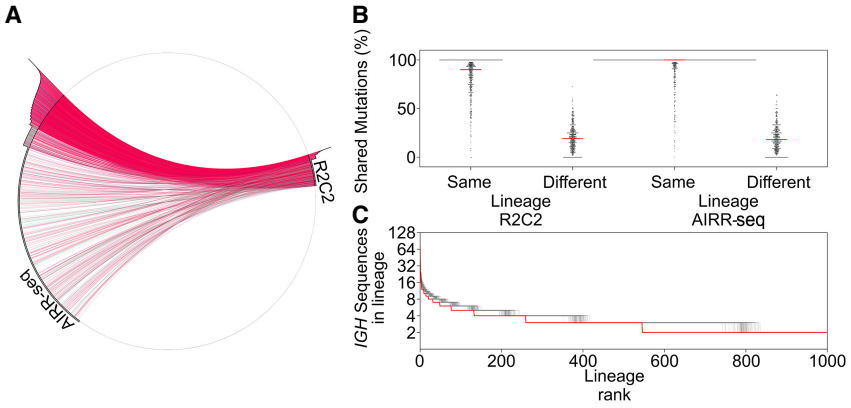


**Figure 6.** Clonal lineages can be measured by R2C2-based repertoires. (*A*) Lineages shared between AIRR-seq and R2C2-based repertoires are indicated by connections *within* a circular plot. Abundance of each lineage within each repertoire is shown as a histogram on the *outside* of the circle. Color of connecting lines and histogram bars indicates the isotype of a lineage: (red) *IGHA*; (black) *IGHM*; (blue) *IGHG*; (green) *IGHD*. (*B*) The distribution of the percentage of mutations in an IGH sequence being shared with IGH sequences within the same or different lineages is shown as swarm plots for R2C2-based and AIRR-seq repertoires. (*C*) The size of lineages ordered by rank is shown for the R2C2-based repertoire (red) and 100 repertoires subsampled from the AIRR-seq repertoire to match the R2C2-based repertoires depth (gray).

enhance specialized assays for the analysis of the complex transcriptomes of human immune cells.

First, we show that R2C2 full-length cDNA data can be used to identify allele-specific expression of isoforms. Identifying allele-specific expression of isoforms represents a formidable challenge for RNA-seq data alone, making long reads basically indispensable for this type of analysis (Deonovic et al. 2017). For our approach, we chose to use highly accurate RNA-seq data to call heterozygous SNPs with the gold-standard GATK workflow (Van der Auwera et al. 2013). This represents a complementary approach to either investigating allele-specific expression in a fully haplotype-resolved sample (GM12878) (Workman et al. 2019) or using long reads themselves to call SNPs (Tilgner et al. 2014). With the accuracy and throughput of our R2C2 method, as well as Pacific Biosciences circular consensus sequencing (PacBio CCS) steadily improving (Wenger et al. 2019), short-read sequencing may soon no longer represent the most accurate way to call SNPs thereby making RNA-seq entirely dispensable for this type of analysis.

Second, we introduce a workflow that can complement short-read RNA-seq data to generate accurate allele-resolved full-length isoforms including the isoforms of all major HLA genes. In the single individual we analyzed, this approach appeared to enable accurate high-resolution HLA typing. However, to establish whether this approach is as reliable and accurate as DNA-based HLA typing will require follow-up studies involving the analysis of many more individuals. We also believe that this approach will be very powerful for the identification of new HLA alleles present in the human population as it does not rely on databases of known HLA alleles as short-read-based HLA typing methods do.

Third and finally, we identified tens of thousands of full-length adaptive immune receptor transcripts in our R2C2 data that can be compiled into repertoires containing a plethora of valuable data about the state of the adaptive immune system (Weinstein et al. 2009). R2C2-based repertoires therefore represent a convenient alternative to specialized AIRR-seq assays for the generation of AIRR data that has been used in conjunction with specialized software (Vander Heiden et al. 2014; Ralph and Matsen 2016) to detect minimal residual disease (Logan et al. 2011) and organ rejection (Vollmers et al. 2015), or for basic research to track B cell clonal lineages or analyze immune aging (de Bourcy et al. 2017a) or class-switching (Horns et al. 2016). In contrast to AIRR-seq methods, generation of R2C2-based repertoires requires no specific primer sets, which makes it a powerful tool for the investigation of not only human but vertebrate adaptive immune receptor diversity.

In summary, R2C2 full-length cDNA is a promising approach for the in-depth analysis of the human immune system and has the potential to replace or enhance specialized RNA-seq, HLA typing, and AIRR-seq approaches in the analysis of clinical samples. Although future studies are still needed to validate this proof-of-concept study as well as establish the scalability of its approach, we hope it provides a stepping-off point for clinical and research assays to leave behind the limitations that short-read RNA-seq imposed on data generation and analysis.

## Methods

### Sample collection and preparation

All experiments were approved by the Internal Review Board at the University of California Santa Cruz. Two whole blood samples were collected from a healthy human adult volunteer by the University of California Santa Cruz Student Health Center ~6 mo apart. Samples were processed by Ficoll gradient (GE Healthcare) to extract PBMCs, which were stored in liquid Nitrogen. RNA was extracted from one PBMC sample using the RNeasy mini kit (Qiagen). DNA was extracted from the other PBMC sample using the MagAttract HMW DNA kit (Qiagen).

### HLA typing

HLA typing was performed at the Immunogenetics and Transplantation Laboratory at University of California, San Francisco.

#### *Sample preparation*

DNA was quantified with NanoDrop (Thermo Fisher Scientific) and adjusted to a concentration of 30 ng/μL. Quality of DNA was assessed by measuring absorbance at $A_{230}$, $A_{260}$, and $A_{280}$. DNA samples were amplified by long-range PCR using the Omixon Holotype HLA genotyping kit, generating full-length gene amplicons for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1* loci. Following PCR, amplicons were cleaned with Exo-SAP (Affymetrix), quantified with QuantiFluor dsDNA system (Promega), and normalized to ~70 ng/μL.

#### *Library preparation and sequencing*

Sequencing libraries were generated for each sample using the Omixon Holotype HLA genotyping kit (Omixon, Inc.). In brief, libraries from individual HLA amplicons were prepared by enzymatic fragmentation, end repair, adenylation, and ligation of indexed adapters. The indexed libraries were pooled and concentrated with Ampure XP beads (Beckman Coulter) before fragment size selection using a PippinPrep (Sage Science), selecting a range of fragments between 650 and 1300 bp. The size-selected library pool was quantified by quantitative PCR (qPCR; Kapa Biosystems) and adjusted to 2 nmol/L. The library was then denatured with NaOH and diluted to a final concentration of 8 pmol/L for optimal cluster density and 600 μL was loaded into the MiSeq reagent cartridge (v2 500 cycle kit). The reagent cartridge and flow cell were placed on the Illumina MiSeq (Illumina) for cluster generation and 2 × 250-bp paired-end sequencing. Samples were demultiplexed on the instrument and the resulting FASTQ files were used for further analysis. HLA genotyping was assigned using Twin version 2.0.1 (Omixon, Inc.) and IMGT/HLA database version 3.24.0_2, using 16,000 read pairs.

### Transcriptome sequencing library preparation

Approximately 200 ng of total RNA was used to generate full-length cDNA using a modified Smart-seq2 protocol (Volden et al. 2018). In short, RNA was reverse transcribed using SMARTscribe RT (Clontech) and ISPCR-Oligo(dT) primer ISPCR-TSO (Supplemental Table S1). Remaining RNA and primer dimers were digested and cDNA was PCR amplified using RNase A, Lambda Exonuclease (NEB), and Kapa Biosystems HiFi HotStart ReadyMix (2X) (KAPA) with the following heat-cycling protocol: for 30 min at 37°C, for 30 sec at 95°C, followed by 12 cycles of (20 sec 98°C; 15 sec 67°C; 10 min 72°C). The reaction was then purified using SPRI beads at a 0.85:1 ratio and eluted in $H_2O$. The resulting full-length cDNA was then used as input into Smart-seq2, R2C2, and Smar2C2 library preparation protocols.

### *Smart-seq2*

Full-length cDNA was then tagmented with Tn5 enzyme (Picelli et al. 2014a) custom loaded with Tn5ME-A/R and Tn5ME-B/R adapters. The Tn5 reaction was performed using 50 ng of cDNA in 5 µL, 1 µL of the loaded Tn5 enzyme, 10 µL of $H_2O$, and 4 µL of 5× TAPS-PEG buffer and incubated for 5 min at 55°C. The Tn5 reaction was then inactivated by the addition of 5 µL of 0.2% sodium dodecyl sulphate and 5 µL of the product was then nick-translated for 6 min at 72°C and further amplified using KAPA HiFi Polymerase (KAPA) using Nextera_Primer_A and Nextera_Primer_B (Supplemental Table S1) with an incubation of 30 sec at 98°C, followed by 13 cycles of (for 10 sec at 98°C, for 30 sec at 63°C, for 2 min at 72°C) with a final extension for 5 min at 72°C. The resulting Illumina library was size selected on an agarose gel to be within 200–400 bp and sequenced on an Illumina NextSeq 2 × 150 run.

### *R2C2*

#### *Splint generation*

To generate the splint, 23 µL of $H_2O$, 25 µL of Kapa Biosystems HiFi HotStart ReadyMix (2×) (KAPA), 1 µL of UMI_Splint_Forward (100 µM), and 1 µL of UMI_Splint_Reverse (100 µM) were incubated for 3 min at 95°C, for 1 min at 98°C, for 1 min at 62°C, and for 6 min at 72°C. The DNA splint was then purified with the Select-a-size DNA Clean and Concentrator kit (Zymo) with 85 µL of 100% EtOH in 500 µL of DNA binding buffer.

#### *Circularization of cDNA*

Next, 200 ng of cDNA was mixed with 200 ng of DNA splint and 2× NEBuilder HiFi DNA Assembly Master Mix (NEB) was added at the appropriate volume. This mix was incubated for 60 min at 50°C. To this reaction we added 5 µL of NEBuffer 2, 3 µL Exonuclease I, 3 µL of Exonuclease III, and 3 µL of Lambda Exonuclease (all NEB), and adjusted the volume to 50 µL using $H_2O$. This reaction was then incubated for 16 h at 37°C followed by a heat inactivation step for 20 min at 80°C. Circularized DNA was then extracted using SPRI beads with a size cutoff to eliminate DNA <500 bp (0.85 beads:1 sample) and eluted in 40 µL of ultrapure $H_2O$.

#### *Rolling circle amplification*

Circularized DNA was split into four aliquots of 10 µL, and each aliquot was amplified in its own 50-µL reaction containing Phi29 polymerase (NEB) and exonuclease resistant random hexamers (Thermo Fisher Scientific) [5 µL of 10× Phi29 Buffer, 2.5 µL of 10 mM (each) dNTPs, 2.5 µL random hexamers (10 µM), 10 µL of DNA, 29 µL ultrapure water, 1 µL of Phi29]. Reactions were incubated overnight at 30°C. T7 Endonuclease was added to each reaction and then incubated for 2 h at 37°C with occasional agitation. The debranched DNA was then extracted using SPRI beads at a 0.5:1 ratio and eluted in 50 µL of $H_2O$.

#### *Oxford Nanopore Technologies sequencing*

The resulting DNA was sequenced across four separate ONT MinION 9.4.1 flow cells. For each run, 1 µg of DNA was prepared using the LSK-109 kit according to the manufacturer's instructions with only minor modifications. End-repair and A-tailing steps were both extend from 5 min to 30 min. The final ligation step was also extended to 30 min. Each run took 48 h and the resulting data in Fast5 format were base called using the high accuracy model of the GPU accelerated Guppy algorithm (version 2.3.5+53a111f, config file: dna_r9.4.1_450bps_flipflop.cfg). To generate R2C2 consensus reads, the resulting raw reads were processed using our C3POa pipeline (https://github.com/rvolden/C3POa).

### *Smar2C2*

Library prep for this protocol is highly similar to Smart-seq2, however instead of cDNA, it uses the debranched rolling circle amplified DNA that is composed of cDNA concatemers. Fifty nanograms of this DNA was tagmented with Tn5 enzyme (Picelli et al. 2014a) custom loaded with Tn5ME-A/R and Tn5ME-B/R adapters. The Tn5 reaction was performed using 50 ng of cDNA in 5 µL, 1 µL of the loaded Tn5 enzyme, 10 µL of $H_2O$ and 4 µL of 5× TAPS-PEG buffer and incubated for 5 min at 55°C. The Tn5 reaction was then inactivated by the addition of 5 µL of 0.2% sodium dodecyl sulphate, and 5 µL of the product was then nick-translated for 6 min at 72°C and further amplified using KAPA HiFi Polymerase (KAPA) using Nextera_Primer_A and Nextera_Primer_B (Supplemental Table S1) with an incubation for 30 sec at 98°C, followed by 13 cycles of (10 sec at 98°C, 30 sec at 63°C, 2 min at 72°C) with a final extension for 5 min at 72°C. The resulting Illumina library was size selected on an agarose gel to be within 200–400 bp and sequenced on an Illumina NextSeq 2 × 150 run.

### *AIRR-seq*

Two hundred nanograms of total RNA was used for cDNA SMARTscribe (Clontech) first-strand synthesis using a primer pool specific to the first exon of all *IGH* isotypes (*IGHM*, *IGHD*, *IGHG1-4*, *IGHA1-2*, *IGHE*) (Supplemental Table S1). In a two-cycle PCR reaction, second and third cDNA strands were synthesized using Kapa Biosystems HiFi HotStart ReadyMix (2×) and two modified primer pools complementary to the beginning of the Framing region 1 (FR1) of the V segment and ~100 bp into the first exon of all IGH isotypes. All primers used in this two-cycle PCR reaction were modified to have unique molecular identifiers and partial Nextera sequences on their 5′ end. cDNA was purified and size selected to >300 nt using Select-a-size DNA Clean and Concentrator kits (Zymo). In a 20-cycle PCR reaction, the cDNA is then amplified with primer completing Nextera sequences as well as Illumina i5 and i7 indexes to enable multiplexing of the libraries. Libraries were then sequenced on the Illumina MiSeq using a 2 × 300 run.

## Data analysis

### *Gene expression*

R2C2 reads were aligned to the hg38 version of the human genome using minimap2 (Li 2018) using "-ax splice --secondary=no" flags and other standard settings. Smart-seq2 and Smar2C2 reads were aligned to the same genome sequence using STAR (version 2.7.1a) (Dobin et al. 2013) and an index built using the GENCODE v29 annotation GTF file. Read alignments were converted in gene expression counts using featureCounts (Liao et al. 2014).

### *Identifying Smar2C2 reads that contain transcript ends*

To identify Smar2C2 reads covering transcript 5′ ends (TSO reads), we determined whether a read contained the sequence of the template switch oligo, which, absent premature template switching, should indicate the 5′ end of a transcript.

To identify Smar2C2 reads covering transcript 3′ ends (Oligo [dT] reads) we analyzed read pairs. First, we determined whether one read of a pair contained the sequence of the Oligo(dT) primer including a stretch of Ts making the other read of that pair a

candidate. We then determine whether a candidate read contained a stretch of Ts which, absent mispriming of the Oligo(dT) primer, should indicate the 3′ end of a transcript.

### *Isoform analysis*

R2C2 reads were analyzed to identify isoforms using version 3 of Mandalorion (Byrne et al. 2017; Volden et al. 2018) and standard settings. Isoforms were categorized using the sqanti_qc.py script of the SQANTI (Tardaguila et al. 2018) program with slight modifications to make it compatible with Python3 and using the GENCODE (v29) annotation of the human genome. Isoform features were extracted from the categorized isoforms using the *ProcessSqantiClassification.py* (Mandalorion utility) script as follows. To identify transcript features not matching the GENCODE annotation, we relied on SQANTI classification:

1. Nonmatching 5′ ends and 3′ ends were identified in isoforms classified as "full-splice_match," "novel_in_catalog," "and novel_not_in_catalog" and had to be located more than 10 nt away from an annotated TSS or poly(A) site.
2. Nonmatching exons were identified from isoforms classified as "novel_not_in_catalog" and had to have no overlap with annotated exons.

### *Allele-specific isoforms*

SNPs present in the sample donor's genome were identified using RNA-seq (Smart-seq2 + Smar2C2) read alignments and GATK (version 3.8-1-0-gf15c1c3ef) following the standard RNA-seq SNP identification workflow (https://software.broadinstitute.org/gatk/documentation/article.php?id=3891). Heterozygous SNPs were phased using R2C2 reads and the new Mandalorion utility TurboPhaser.py, taking advantage of R2C2 reads spanning entire gene loci and grouping SNPs that appeared in the same reads. TurboPhaser.py also sorted R2C2 reads and RNA-seq reads into alleles based on the SNPs they contained. The sorted R2C2 reads were then used in the Mandalorion pipeline to identify isoform sequences. The sequences were then error-corrected using Pilon (Walker et al. 2014) and RNA-seq reads aligned to the isoform sequences using minimap2 with the "-x sr" preset.

### *HLA typing*

Using the HLAtyping.py script allele-specific error-corrected isoform sequences were aligned to the human genome using minimap2 ("--secondary=no -x splice"). Isoforms aligning to HLA loci were then realigned to HLA transcript sequences retrieved from the IPD-IMGT/HLA database (Robinson et al. 2015, 2000) using minimap2 ("-ax splice -N 100"). For each HLA gene and allele, the best full-length match was reported. For RNA-seq-only approaches, Smart-seq2 reads and Smar2C2 reads were pooled and processed as required by seq2HLA and arcasHLA and the programs were run with standard settings.

### *AIRR analysis*

R2C2 reads aligning to adaptive immune receptor loci were extracted using SAMtools (Li et al. 2009). The sequences were then analyzed using IgBLAST (Ye et al. 2013) with V, D, and J segments retrieved from IMGT (Lefranc et al. 2004). For the in-depth analysis of BCR IGH repertoires, this output was then parsed using custom scripts to report CDR3 length and V segment, as well as the positions of mismatches in the V segments. Isotype and isoform (secreted vs. membrane-bound) of each sequence were determined by comparing the part of extracted R2C2 reads corresponding to the constant region to a database of isotype and isoform sequences and, if it was high quality enough, the best match was used. Sequences were then grouped into lineages using a simple single linkage clustering approach using a 90% CDR3 nucleotide similarity cutoff.

### Code access

Mandalorion and its utilities for isoform identification and sequence determination are available on GitHub (https://github.com/rvolden/Mandalorion-Episode-III). The Mandalorion package also contains scripts for processing SQANTI classification, sorting R2C2 reads into alleles, and HLA typing.

C3POa is available on GitHub (https://github.com/rvolden/C3POa). The C3POa GitHub also contains scripts for identifying and merging reads with matching UMIs. Scripts for the parsing and grouping AIRR data into lineages is available at GitHub (https://github.com/christopher-vollmers/AIRR). Analysis scripts not published previously are also available as Supplemental Scripts.

### Data visualization

All data visualization was done using Python/Numpy/Scipy/Matplotlib (https://www.scipy.org) (Hunter 2007; Oliphant 2007; van der Walt et al. 2011). Schematics were drawn in Inkscape (https://inkscape.org/en/).

## Data access

The R2C2 and RNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA559668. The TMI-seq generated for a previous study have been submitted to the NCBI BioProject database under accession number PRJNA559668. Processed data are available at https://users.soe.ucsc.edu/~vollmers/PBMC_data/R2C2_reads.fa and https://users.soe.ucsc.edu/~vollmers/PBMC_data/AIRRseq_reads.fa.

## Competing interest statement

C.V. and R.V. have filed patent applications on aspects of the R2C2 method and data analysis used in the manuscript.

## Acknowledgments

We thank Dr. Raja Rajalingam and the Immunogenetics and Transplantation Laboratory (ITL) at the University of California, San Francisco, for performing high-resolution HLA typing. We thank Ed Malloy, Aisha Coons, and Jerrold Michaud of the Student Health Center at the University of California Santa Cruz for expert assistance with blood draws. We acknowledge funding by the National Human Genome Research Institute/National Institute of Health Training Grant 1T32HG008345-01 (to C.C., A.B., and R.V.), the Hellman Foundation, Santa Cruz Cancer Benefit Group, and National Institute of General Medical Sciences/National Institute of Health Grant 1R35GM133569-01 (to C.V.).

*Author contributions:* C.C., A.B, and M.A. performed experiments. C.C., R.V., and C.V. analyzed the data. C.C. and C.V. conceived of the study and designed experiments. C.V., C.C., A.B., R.V., and M.A. wrote and edited the manuscript.

## References

Anders S, Huber W. 2012. *Differential expression of RNA-Seq data at the gene level–the DESeq package*. European Molecular Biology Laboratory

(EMBL), Heidelberg, Germany. http://www.genomatix.de/online_help/help_regionminer/DESeq_1.10.1.pdf.
Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, Lee JC, Thomas DC, Flint SM, Kellam P, Jayne DRW, et al. 2019. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574:** 122–126. doi:10.1038/s41586-019-1595-3
Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, Türeci O, Diken M, Castle JC, Sahin U. 2012. HLA typing from RNA-Seq sequence reads. *Genome Med* **4:** 102. doi:10.1186/gm403
Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, Hemmers S, Putintseva EV, Obraztsova AS, Shugay M, et al. 2017. Antigen receptor repertoire profiling from RNA-seq data. *Nat Biotechnol* **35:** 908–911. doi:10.1038/nbt.3979
Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun* **8:** 16027. doi:10.1038/ncomms16027
Byrne A, Supple MA, Volden R, Laidre KL, Shapiro B, Vollmers C. 2019. Depletion of hemoglobin transcripts and long-read sequencing improves the transcriptome annotation of the polar bear (*Ursus maritimus*). *Front Genet* **10:** 643. doi:10.3389/fgene.2019.00643
Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. 2016. Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier–guided amplicon assembly. *J Immunol* **196:** 2902–2907. doi:10.4049/jimmunol.1502563
Davila ML, Brentjens RJ. 2016. CD19-targeted CAR T cells as novel cancer immunotherapy for relapsed or refractory B-cell acute lymphoblastic leukemia. *Clin Adv Hematol Oncol* **14:** 802–808.
de Bourcy CFA, Angel CJL, Vollmers C, Dekker CL, Davis MM, Quake SR. 2017a. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc Natl Acad Sci* **114:** 1105–1110. doi:10.1073/pnas.1617959114
de Bourcy CFA, Dekker CL, Davis MM, Nicolls MR, Quake SR. 2017b. Dynamics of the human antibody repertoire after B cell depletion in systemic sclerosis. *Sci Immunol* **2:** eaan8289. doi:10.1126/sciimmunol.aan8289
Deonovic B, Wang Y, Weirather J, Wang XJ, Au KF. 2017. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Res* **45:** e32. doi:10.1093/nar/gkw1076
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635
Fischer J, Paret C, El Malki K, Alt F, Wingerter A, Neu MA, Kron B, Russo A, Lehmann N, Roth L, et al. 2017. *CD19* isoforms enabling resistance to CART-19 immunotherapy are expressed in B-ALL patients at initial diagnosis. *J Immunother* **40:** 187–195. doi:10.1097/CJI.0000000000000169
Fry TJ, Shah NN, Orentas RJ, Stetler-Stevenson M, Yuan CM, Ramakrishna S, Wolters P, Martin S, Delbrook C, Yates B, et al. 2018. CD22-targeted CAR T cells induce remission in B-ALL that is naive or resistant to CD19-targeted CAR immunotherapy. *Nat Med* **24:** 20–28. doi:10.1038/nm.4441
Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31:** 3356–3358. doi:10.1093/bioinformatics/btv359
Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36:** 1197–1202. doi:10.1038/nbt.4259
Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111
Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, Davis MM, Quake SR. 2016. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *eLife* **5:** e16578. doi:10.7554/eLife.16578
Horns F, Vollmers C, Dekker CL, Quake SR. 2019. Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. *Proc Natl Acad Sci* **116:** 1261–1266. doi:10.1073/pnas.1814213116
Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng* **9:** 90–95. doi:10.1109/MCSE.2007.55
Kulkarni S, Ramsuran V, Rucevic M, Lied A, Kulkarni V, Le Gall S, Carrington M. 2017a. Alternative polyadenylation signals regulate HLA-A surface expression. *J Immunol* **198:** 124.16–124.16.
Kulkarni S, Ramsuran V, Rucevic M, Singh S, Lied A, Kulkarni V, O'hUigin C, Le Gall S, Carrington M. 2017b. Posttranscriptional regulation of HLA-A protein expression by alternative polyadenylation signals involving the RNA-binding protein syncrip. *J Immunol* **199:** 3892–3899. doi:10.4049/jimmunol.1700697
Lefranc MP, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, et al. 2004. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol* **4:** 17–29.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352
Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656
Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buno I, Armstrong R, Fire AZ, Weinberg KI, et al. 2011. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc Natl Acad* **108:** 21194–21199. doi:10.1073/pnas.1118357109
Mose LE, Selitsky SR, Bixby LM, Marron DL, Iglesia MD, Serody JS, Perou CM, Vincent BG, Parker JS. 2016. Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics* **32:** 3729–3734. doi:10.1093/bioinformatics/btw526
Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. 2016. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep* **6:** 31602. doi:10.1038/srep31602
Oliphant TE. 2007. Python for scientific computing. *Comput Sci Eng* **9:** 10–20. doi:10.1109/MCSE.2007.58
Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R. 2019. arcasHLA: high resolution HLA typing from RNAseq. *Bioinformatics* **36:** 33–40. doi:10.1093/bioinformatics/btz474
Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014a. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* **24:** 2033–2040. doi:10.1101/gr.177881.114
Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014b. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9:** 171–181. doi:10.1038/nprot.2014.006
Ralph DK, Matsen FA 4th. 2016. Likelihood-based inference of B cell clonal families. *PLoS Comput Biol* **12:** e1005086. doi:10.1371/journal.pcbi.1005086
Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE. 2000. IMGT/HLA Database—a sequence database for the human major histocompatibility complex. *Tissue Antigens* **55:** 280–287. doi:10.1034/j.1399-0039.2000.550314.x
Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* **43:** D423–D431. doi:10.1093/nar/gku1161
Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TYK, Pietzsch J, Fenyo D, Abadir A, Velinzon K, et al. 2011. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333:** 1633–1637. doi:10.1126/science.1207227
Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31:** 1009–1014. doi:10.1038/nbt.2705
Shiina T, Hosomichi K, Inoko H, Kulski JK. 2009. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet* **54:** 15–39. doi:10.1038/jhg.2008.5
Sotillo E, Barrett DM, Black KL, Bagashev A, Oldridge D, Wu G, Sussman R, Lanauze C, Ruella M, Gazzara MR, et al. 2015. Convergence of acquired mutations and alternative splicing of *CD19* enables resistance to CART-19 immunotherapy. *Cancer Discov* **5:** 1282–1295. doi:10.1158/2159-8290.CD-15-1020
Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11:** 1438. doi:10.1038/s41467-020-15171-6
Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28:** 396–411. doi:10.1101/gr.222976.117
Tilgner H, Grubert F, Sharon D, Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci* **111:** 9869–9874. doi:10.1073/pnas.1400447111
Tonegawa S. 1983. Somatic generation of antibody diversity. *Nature* **302:** 575–581. doi:10.1038/302575a0

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43:** 11.10.1–11.10.33. doi:10.1002/0471250953.bi1110s43

Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, Vigneault F, Kleinstein SH. 2014. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30:** 1930–1932. doi:10.1093/bioinformatics/btu138

van der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* **13:** 22–30. doi:10.1109/MCSE.2011.37

Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci* **115:** 9726–9731. doi:10.1073/pnas.1806447115

Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci* **110:** 13463–13468. doi:10.1073/pnas.1312146110

Vollmers C, De Vlaminck I, Valantine HA, Penland L, Luikart H, Strehl C, Cohen G, Khush KK, Quake SR. 2015. Monitoring pharmacologically induced immunosuppression by immune repertoire sequencing to detect acute allograft rejection in heart transplant patients: a proof-of-concept diagnostic accuracy study. *PLoS Med* **12:** e1001890. doi:10.1371/journal.pmed.1001890

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9:** e112963. doi:10.1371/journal.pone.0112963

Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Viña MA, Davis RW, Davis MM, et al. 2012. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proc Natl Acad Sci* **109:** 8676–8681. doi:10.1073/pnas.1206614109

Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324:** 807–810. doi:10.1126/science.1170020

Weng WK, Armstrong R, Arai S, Desmarais C, Hoppe R, Kim YH. 2013. Minimal residual disease monitoring with high-throughput sequencing of T cell receptors in cutaneous T cell lymphoma. *Sci Transl Med* **5:** 214ra171. doi:10.1126/scitranslmed.3007420

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16:** 1297–1305. doi:10.1038/s41592-019-0617-2

Ye J, Ma N, Madden TL, Ostell JM. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* **41:** W34–W40. doi:10.1093/nar/gkt382

*Received September 14, 2019; accepted in revised form April 3, 2020.*