University of California
Santa Cruz

Single-Cell Nanopore Sequencing

A dissertation submitted in partial satisfaction of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Roger S. Volden

June 2021

The Dissertation of Roger S. Volden is
approved:

_____
Professor Christopher Vollmers, Chair

_____
Professor Angela Brooks

_____
Professor Richard Edward Green

_____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

Table of figures

Single-Cell Nanopore Sequencing
Roger S. Volden
Abstract

Most transcriptomic analyses are done using Illumina short-read sequencing. While these analyses can be used for highly accurate annotation of individual splice junctions, they are incapable of piecing together combinations of splice junctions to reveal complete RNA transcript isoforms. Exon connectivity information is required for accurate full-length RNA transcript isoform analyses. While long-read sequencing technologies like Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) can provide exon connectivity information, neither provide a cost effective way to produce high accuracy full-length reads. I present the ONT-based Rolling Circle Amplification to Concatemeric Consensus (R2C2) and Concatemeric Consensus Caller with Partial Order alignments (C3POa) methods, which generate more accurate reads of full-length RNA transcript isoforms than other long-read sequencing methods. I apply these methods to full-length RNA isoform sequencing in single-cells for differential isoform expression across cell types.

Acknowledgements

The text of this dissertation includes reprints of the following previously published material: Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA (Volden et al. 2018), Realizing the potential of full-length transcriptome sequencing (Byrne et al. 2019), and Highly Multiplexed Single-Cell Full-Length cDNA Sequencing of human immune cells with 10X Genomics and R2C2 (Volden et al. 2021). The co-authors listed in these publications directed and supervised the research which forms the basis for the dissertation.

I would like to thank my family for encouraging me to study science and constantly pushing me to be the best that I can be. My parents and sister are an excellent support system and I am very lucky to have people that care about me so much. I would also like to thank my lab mates and mentors, Ashley Byrne and Charles Cole. As a budding scientist, I always looked up to you guys. Ashley is one of the best molecular biologists I know and Charles is an amazing bioinformatician. I really look forward to seeing what you both do in the future.

Lastly, I would like to thank my good friend and mentor, Chris Vollmers. His passion for science is infectious and I would not be the scientist or person that I am today without him. He has taught me countless things over the years in both science and in life. I truly believe that we should all strive to be more like him. He is, without a doubt, the best mentor that anyone could ever wish for and I will miss working with him dearly.

Single-Cell Nanopore Sequencing

cDNA Sequencing

Parts of this section are adapted from **Realizing the potential of full-length transcriptome sequencing** (Bryne et al. 2019).

A Brief History of RNA and cDNA sequencing

The analysis of what RNA transcripts (annotation) are present in a sample and at what level (quantification) has relied on a mix of technologies over the last three decades. Early efforts to annotate and quantify complex eukaryotic transcriptomes were highly labor intensive. During the early 1990's, efforts to evaluate RNA sequences on a large scale relied heavily on ESTs (Expressed Sequence Tags) whereby cDNA molecules were individually cloned, screened, and Sanger-sequenced to determine full-length mRNA sequences and observe semi-quantitative changes in gene expression[1]. The Sanger-sequencing based SAGE (Serial Analysis of Gene Expression) method improved quantification and reduced cost by concatenating smaller 15-20 bp fragments of many cDNA molecules together for sequencing[2]. However, because of the short length of analyzed fragments SAGE was inherently less useful for annotation. Hybridization-based microarray approaches completely eschewed annotation but simplified the quantification of already annotated genes[3].

The introduction of massively parallel sequencing in the mid-to-late 2000s completely changed transcriptome annotation and quantification. When massively parallel sequencing – best represented by the now dominant Illumina technology – became available to research labs it could generate millions of sequencing reads at a length of ~30 nucleotides (nt). Although initially intended for the sequencing of genomic DNA, researchers quickly found ways to leverage the power of these sequencers for transcriptome analysis in the form of the RNA-seq assay. RNA-seq sequences short cDNA fragments at extremely high throughput and quickly displaced microarray-based transcriptome analysis for a number of reasons

including cost considerations as well as the ability to detect previously unknown transcripts and quantify the use of individual splice sites. In the last decade, with a few hiccups , Illumina sequencers have steadily and massively improved, although these improvements have come with compromises in experimental design. Most prominently newer Illumina sequencers require additional precautions to avoid sample cross-contamination during the sequencing reaction.

Current Illumina sequencers like the NovaSeq can generate billions of sequencing reads at a length of 150 nt allowing the multiplexed analysis of hundreds to thousands of samples in a single run. At this read-length and output, RNA-seq reads are not only useful for transcriptome quantification but also for annotation. Consequently, efforts like GENCODE and RefSeq heavily rely on this data type for their respective annotation approaches[4,5]. Paired with literally hundreds of sample preparation techniques and analysis pipelines, transcriptome analysis by short-read RNA-seq[6] is now a core component of research in nearly all fields of biology.

So, while it is clear that RNA-seq has revolutionized transcriptome annotation and quantification it is also becoming increasingly clear that it is ultimately a stop-gap solution of limited power born out the limitations of short-read sequencing. These limitations prevent RNA-seq from annotating and quantifying transcriptomes on the level or RNA transcript isoforms, i.e. transcript variants expressed by the same gene utilizing combinations of alternative splice sites, transcription start sites, and transcription termination or polyA sites. Thus, to fully understand the fundamentals of gene expression, isoform information will be required.

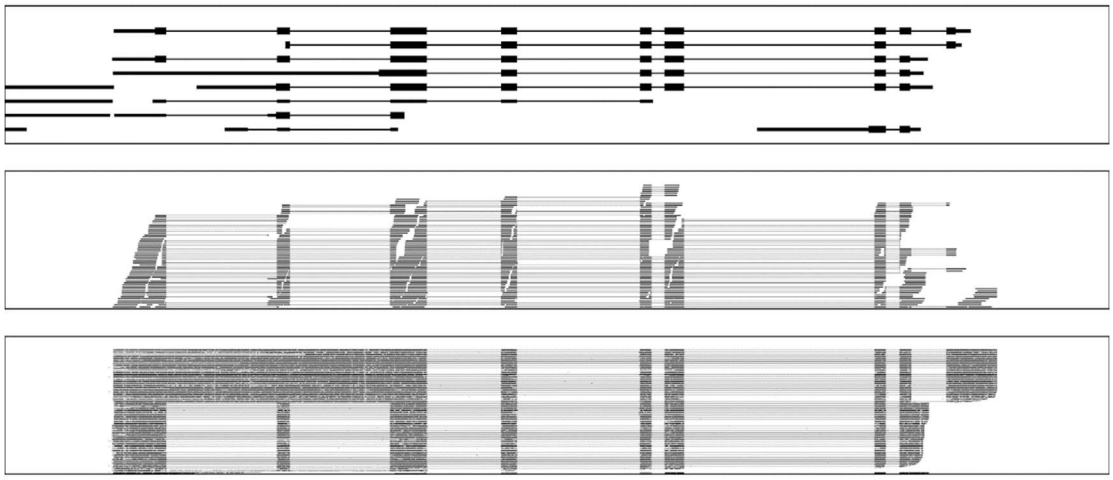

Fig. 1: Fundamental difference between short- and long-read sequencing of transcripts. *Short RNA-seq reads (middle) only capture small fragments of transcripts. RNA-seq data therefore lacks unambiguous isoform data (top) leading to the inference of many erroneous isoforms. Long-read full-length cDNA data (bottom) captures transcripts end-to-end making isoform inference unambiguous.*

Despite its dominant position in transcriptome analysis, short-read RNA-seq has so far failed at capturing the true complexity of eukaryotic transcriptomes. While RNA-seq can interrogate individual transcript features like splice sites, transcription start sites, and polyA sites, it fails at determining how these individual features are combined into transcript isoforms. This is due to the fact that the read length of short-read sequencers is too short to capture entire transcripts from end-to-end (Fig. 1). Incomplete fragments of transcripts therefore have to be computationally assembled into full-length isoforms. This is done using powerful algorithms performing de-novo (e.g. Trinity[7], rnaSPAdes[8]) or genome-guided transcriptome assemblies (e.g. Cufflinks[9], StringTie[10]). All of these assemblers ultimately fail at discerning complex transcript isoforms expressed by the same gene because of limitations of the underlying data. First, RNA-seq reads often do not cover the ends of transcripts leaving TSS and polyA sites unresolved[11]. Second, alternative transcript features are too far apart to be resolved by RNA-seq raw data, i.e. if a transcript has two alternative splice sites 1000bp apart, no individual RNA-seq read will ever connect those two events.

Computational methods that take this into account have been developed, however they still fail at deconvoluting complex isoform mixtures[12].

cDNA Isoform Sequencing with Long Reads

RNA-seq cannot analyze full-length cDNA because it is limited in the length of the sequences it can process. Third generation sequencing technologies like Oxford Nanopore Technologies (ONT) nanopore based sequencing and Pacific Biosciences (PacBio) zero mode waveguide (ZMW) based sequencing overcome this length limitation, thereby enabling the sequencing of entire cDNA molecules.

Sequencing entire cDNA molecules allows long read technologies to define and quantify isoforms which is largely impossible using short reads[13] (Fig. 1). Long reads allow you to encompass full isoforms, thereby eliminating error-prone inference based on fragmented short-read data when assembling isoforms[14,15] (Fig. 1). Just like short-read sequencing, long-read technology was initially intended for genomic DNA sequencing, but it was only a matter of time until cDNA copies of RNA transcript molecules were sequenced on PacBio and ONT sequencers.

Initial studies used long reads for the targeted analysis of specific highly complex transcripts[16] or to add small amounts of long-read data to short-read RNA-seq data[17,18]. Increasing read throughput has allowed the analysis of whole transcriptomes of diverse organisms with long-read data alone[15,19–21] and in addition to the analysis of cDNA, ONT sequencers now offer the ability to sequence RNA directly[22,23]. Finally, long-read technology has been used to analyze the transcriptomes of single cells[24–26].

These papers clearly highlight the potential of long-read sequencing to identify new isoforms and isoform features like new splice sites, TSSs, and polyA sites which is essential to unambiguously annotate and quantify transcriptomes. These papers also lay out a path for the future: In the short-term, long-read technology will be a boon for the transcriptome annotation of non-model organisms. With a moderate investment generating long-read

transcriptome data for a variety of tissues and organs present in a non-model organism, transcriptome annotations will get close to the comprehensiveness and quality of highly curated mouse and human transcriptomes. In the long-term, we believe long-read technology has the potential to entirely replace short-read RNA-seq for transcriptome analysis.

Single-cell cDNA Sequencing

In addition to development in sequencing technology and analysis, a large proportion of technology development in the last 5 years has been focused on enabling the transcriptome analysis of single cells. Traditional cDNA sequencing methods analyze the cDNA of many different cells in bulk, essentially giving an average of a sample or population. Single-cell transcriptomics using short reads is a powerful tool in determining gene expression profiles and identifying individual transcript features. These profiles can be used to computationally sort cells into clusters to identify new cell types and define their transcriptional profiles without any prior knowledge of the population. However, short-read sequencing technologies lose long distance information because of fragmentation of cDNA during the library preparation. Loss of long distance connectivity takes away from the ability to determine isoform-based cell heterogeneity. To regain this lost information and gain further insight into cell heterogeneity, single cell cDNA must be analyzed using long reads. Using standard ONT methods, we have shown that single B cells show high levels of heterogeneity between cells[26].

Single-cell cDNA Sequencing with FACS

Before the advent of droplet-based single-cell sequencing, fluorescence activated cell sorting (FACS) was the most common approach to separate samples into single cells[27]. Compared to other early methods for single-cell sequencing like serial dilution[28], FACS is relatively fast and high throughput. FACS also allows for separation of cells based on various cellular properties (e.g., size, fluorescence, granularity), which simplifies selecting for specific cell types.

The main disadvantages of using FACS when compared to droplets are that the speed and throughput are low, confirmation bias when analyzing certain cell types, and labor intensity. The speed and throughput of FACS is low because single cells need to be sorted into PCR plates, which allows for a practical maximum of 384 cells. While it is possible to use multiple plates, multiplexing these libraries becomes challenging and labor intensive. Another disadvantage of FACS is that when sorting cells, you need to be selecting for or against something (e.g., surface marker, size). This selection has a couple of facets: sorting cells based on prior knowledge creates confirmation bias, and inconsistent gating can exacerbate this bias. For example, it is expected that cells sorted for CD8 would mostly be cytotoxic T cells[29]. However, when looking at the transcriptomic expression values, those cells might not express CD8 very much. This mismatch between transcript expression and what the cell presents on its surface can work for and against FACS. What makes this worse is that setting gates for cell sorting is highly subjective. This can lead to high variability between experiments.

Single-cell cDNA Sequencing with Droplets

While large-scale efforts to analyze single cell transcriptomes have continued to rely on standard PCR plates[30], this approach is very labor intensive and expensive. 10X Genomics has commercialized a droplet-based approach that can enable the analysis of thousands of single cells in a single experiment. 10X Genomics uses a microfluidics-based workflow with microbeads and barcoding oligos for single-cell transcriptome sequencing. Beads that are coated with primers are combined with cells, which are separated by an oil sheath. After the cells and beads are paired, the cells are lysed so they release their RNA for a reverse transcription (RT) reaction. Following the RT, the emulsion that the cells are in is broken and all of the cDNA is amplified. An outline of this process is shown in Figure 2.

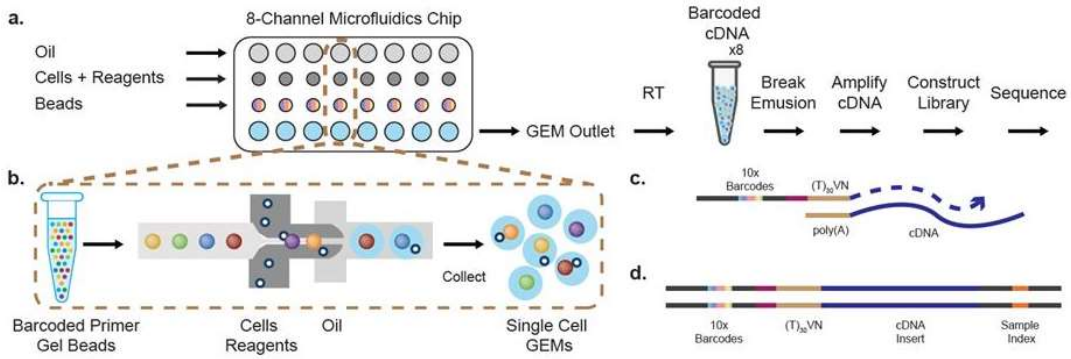

Fig. 2: Overview of the 10X genomics single cell sequencing workflow*. Adapted from DNA Technologies Core.*

This results in a large library of cDNA that contains single-cell barcodes as well as unique molecular identifiers (UMIs) which can be demultiplexed computationally. Most importantly, even though this cDNA is sheared for Illumina analysis, the protocol generates barcoded full-length cDNA as an intermediate product which could be used as input into long-read sequencing assays.

For my graduate work, several steps needed to be taken to apply long reads to highly-multiplexed single-cell sequencing while maintaining cost efficiency. The first step is there needs to be a highly accurate, cheap, and high throughput long-read sequencing method. For my first aim, I introduce my own molecular biology and computational methods to produce highly accurate long reads without compromising on cost or throughput.

Aim 1: Improving nanopore read accuracy with the R2C2 and C3POa methods

This section is adapted from **Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA** (Volden et al. 2018).

R2C2: Improving nanopore read accuracy enables the sequencing of highly-multiplexed full-length single-cell cDNA

Roger Volden[1,3], Theron Palmer[1,3], Ashley Byrne[2,3], Charles Cole[1,3], Robert J Schmitz[4], Richard E Green[1,3], Christopher Vollmers[1,3,*]
1) Department of Biomolecular Engineering, University of California Santa Cruz, CA 95064, USA
2) Department of Molecular, Cellular, Developmental Biology, University of California Santa Cruz, CA 95064, USA
3) UC Santa Cruz Genomics Institute, Santa Cruz, California 95064, USA
4) Department of Genetics, University of Georgia, Athens, GA 30602, USA
*) Corresponding author. Email: vollmers@ucsc.edu

Abstract

High-throughput short-read sequencing has revolutionized how transcriptomes are quantified and annotated. However, while Illumina short-read sequencers can be used to analyze entire transcriptomes down to the level of individual splicing events with great accuracy, they fall short of analyzing how these individual events are combined into complete RNA transcript isoforms. Because of this shortfall, long-read sequencing is required to complement short-read sequencing to analyze transcriptomes on the level of full-length RNA transcript isoforms. However, there are issues with both Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) long-read sequencing technologies that prevent their widespread adoption. Briefly, PacBio sequencers produce low numbers of reads with high accuracy, while ONT sequencers produce higher numbers of reads with lower accuracy. Here we introduce and validate a new long-read ONT based sequencing method. At the same cost, our Rolling Circle Amplification to Contameric Consensus (R2C2) method generates more accurate reads of full-length RNA transcript isoforms than any other available long-read

sequencing method. These reads can then be used to generate isoform-level transcriptomes for both genome annotation and differential expression analysis in bulk or single cell samples.

Introduction

Short-read RNAseq has been used for the analysis of transcriptomes for over a decade[6]. The massive read output of Illumina sequencers makes it possible to quantify gene expression accurately using this approach. However, to accommodate Illumina sequencers' short read-length, RNA or cDNA has to be fragmented during sample preparation, thereby losing long distance RNA transcript isoform information. Specialized protocols like SLR[31] or spISO-seq[32] have been used successfully to recover long-distance information but they require either specialized instrumentation or complex workflows. The SLR method assembles mostly incomplete cDNA molecules, and has limited throughput, while spISO-seq requires a 10X Genomics instrument and generates read clouds which capture long distance information, and yet cannot assemble full-length cDNA molecules.

In contrast, long-read sequencing technology has the capability to sequence entire cDNA molecules end-to-end. Currently, the PacBio Iso-Seq pipeline represents a powerful gold standard for cDNA sequencing[15] and has been used to investigate a wide range of transcriptomes[33,34]. The PacBio Sequel sequencer produces ~200k accurate circular consensus reads of full-length cDNA molecules per run.

ONT technology could present a valuable alternative for cDNA sequencing, because the ONT MinION can currently generate more than one million reads per run. We and others have shown that the ONT MinION can sequence cDNA at high throughput, but that data analysis is challenging[26,35] due to its high error rate. Base level identification of splice junction sequence is the main challenge.

One strategy to increase the base accuracy of cDNA sequences produced by the higher-output ONT MinION sequencer is to apply the circular consensus principle applied by

PacBio sequencers. By sequencing 16S amplicon molecules, the INC-seq[36] method has shown that this is possible, in principle. But, the reported throughput of a few thousand reads per-run would be insufficient for transcriptome analysis. Further, like PacBio technology, the INC-Seq method uses blunt-end ligation to circularize double-stranded DNA molecules, which does not differentiate between full-length or fragmented DNA molecules. In summary, current technology produces reads that are either too inaccurate (ONT), potentially incomplete (Illumina, PacBio, ONT, INC-seq), or too low-throughput/expensive (PacBio, SLR, INC-seq) to enable high-throughput complete cDNA sequencing.

Here we introduce the Rolling Circle to Concatemeric Consensus (R2C2) method which overcomes these limitations by leveraging the long read length of the ONT technology to generate consensus sequences with increased base accuracy. First, we benchmark R2C2 against the PacBio Iso-Seq gold standard for the analysis of the same synthetic transcript mixture. Second, we apply R2C2 to analyze the transcriptomes of 96 single B cells derived from a healthy adult. We show that a single run of R2C2 can generate over 400,000 reads covering full-length cDNA molecules with a median base accuracy of 94%. Using a new version of our Mandalorion pipeline, these reads can be used to identify high confidence RNA transcript isoforms present in bulk or single cell transcriptomes. Illustrating the power of this approach, we find that many of the B cells in our study express RNA transcript isoforms of the CD19 gene that lack the epitope targeted by CAR T-cell therapy[37–39].

Results

R2C2 improves the base accuracy of the ONT MinION.

To benchmark the R2C2 method, we analyzed SIRV E2 synthetic spike-in RNA. First, we reverse transcribed and amplified the synthetic spike-in RNA using the Tn5Prime[40] protocol, which is a modification of the Smart-seq2 protocol which uses a distinct template switch oligo containing 7 nucleotide sample indexes during reverse transcription. Amplification introduces an additional 8 nucleotide index into the cDNA molecules. The amplified cDNA is then

circularized using a DNA splint and the NEBuilder Hifi DNA Assembly Master Mix, a proprietary variant of Gibson Assembly. The DNA splint is designed to circularize only full-length cDNA terminating on both ends in sequences complementary to the primers used to amplify cDNA (Fig. 3). Circularized cDNA is then amplified using Phi29 and random hexamers to perform Rolling Circle Amplification (RCA). The resulting High Molecular Weight (HMW) DNA was then debranched using T7 Endonuclease and sequenced on the ONT MinION sequencer using the 1D sequencing kit (LSK108) kit and R9.5 flowcell (FLO-MIN107).

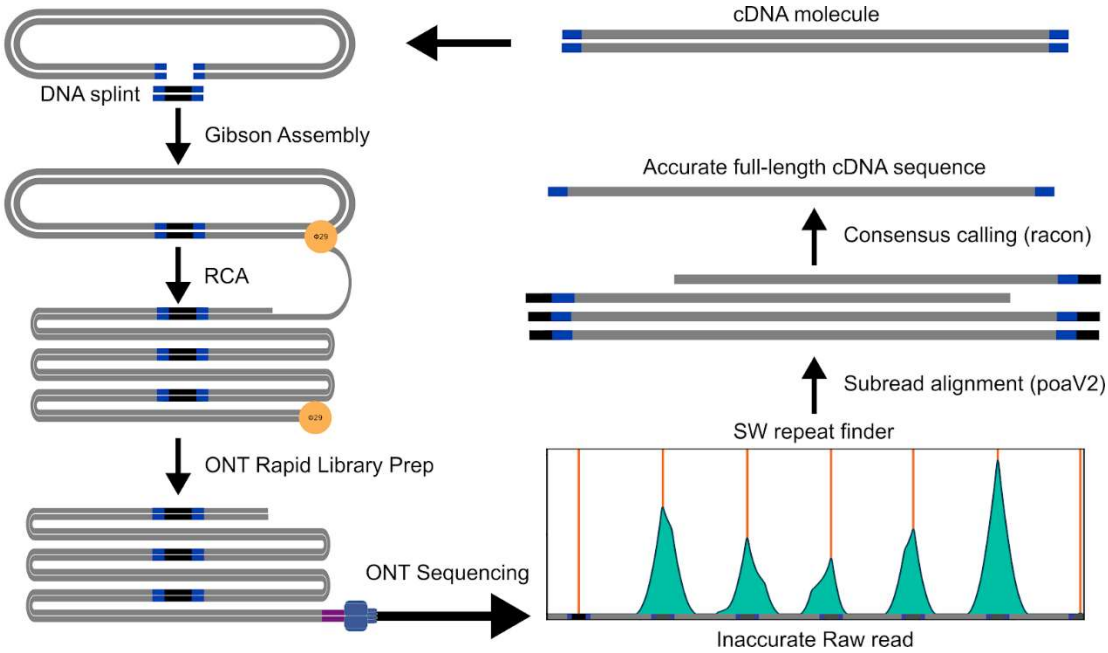

Fig. 3: R2C2 method overview. *cDNA is circularized using Gibson Assembly, amplified using RCA, and sequenced using the ONT MinION. The resulting raw reads are split into subreads containing full-length or partial cDNA sequences, which are combined into an accurate consensus sequences using our C3POa workflow which relies on a custom algorithm to detect DNA splints as well as poaV2 and racon.*

The sequencing run produced 828,684 reads with an average length of 5.0kb resulting in a total base output of 4.1Gb. For downstream analysis we selected 621,970 of these reads that were longer than 1kb and had a raw quality score(Q) ≥ 9. We next used our C3POa

(Contameric Consensus Caller using POA) computational workflow to generate full-length cDNA consensus reads from the raw reads. C3POa detects DNA splint sequences raw reads using BLAT[41]. Because BLAT is likely to miss DNA splint sequences in the noisy raw reads, we analyze each raw read for which BLAT found at least one DNA splint sequence with a custom repeat finder which parses the score matrix of a modified Smith-Waterman self-to-self alignment (Fig 3, Fig. 4A). Repeats, or subreads, are then combined into a consensus and error-corrected using poaV2[42] and racon[43], respectively. Finally, only reads containing known priming sites at both cDNA ends are retained. In this way, C3POa generated 435,074 full-length cDNA consensus reads (and an additional 46,994 consensus reads from another multiplexed experiment) with varying subread coverage (Table 1, Fig. 4A,B).

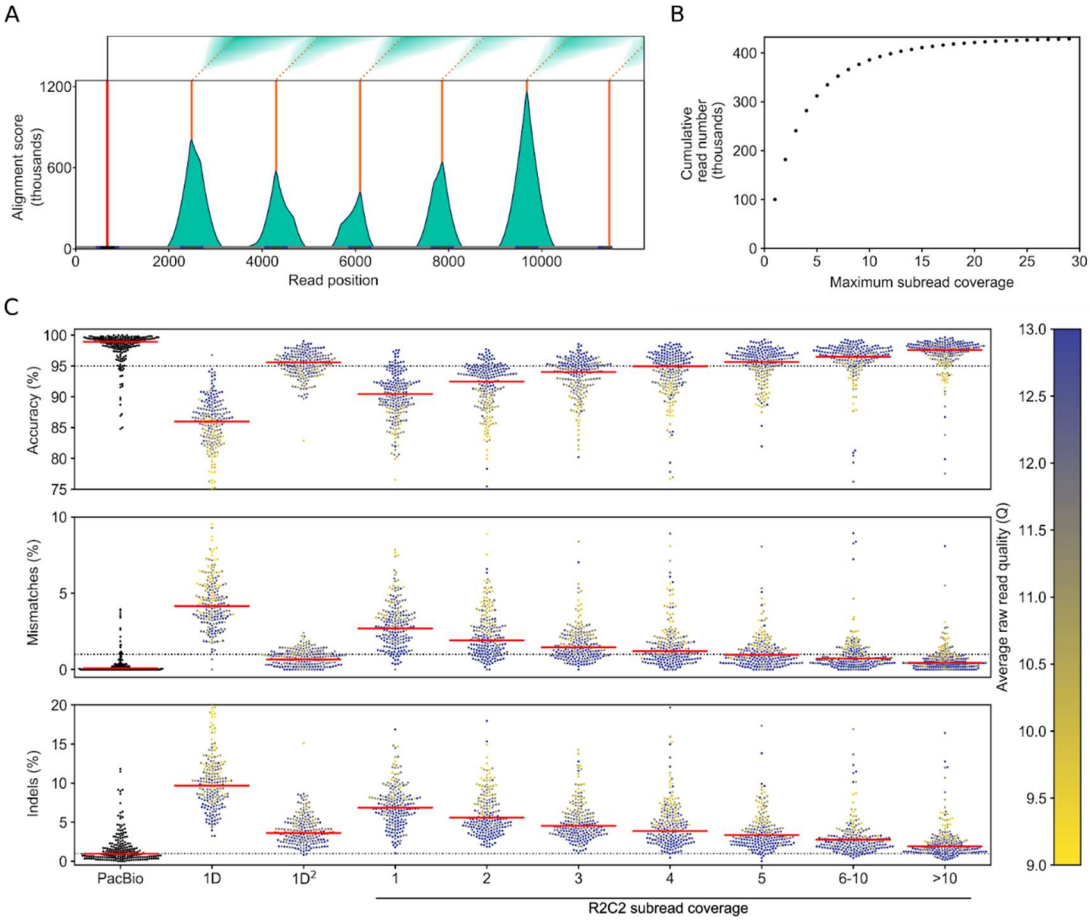

Fig. 4: Raw reads are processed into consensus reads of varying subread coverage

A) Example of a 11.5kb raw ONT read that was analyzed by our custom Smith-Waterman repeat finder. One initial splint (red line) is identified using the BLAT aligner, then modified Smith-Waterman self-to-self alignments are performed starting from the location of the initial splint. The score matrices (on top) are then processed to generate alignment score histograms (teal). We then call peaks (orange) on these histograms. Complete subreads are then defined as the sequences between two peaks. B) Cumulative number of SIRV E2 R2C2 consensus reads is plotted against their subread coverage. C) PacBio Isoseq, standard ONT 1D, and 1D² are compared to R2C2 at different subread coverage. Read accuracy is determined by minimap2 *alignments to SIRV transcripts (see Methods). Median accuracy is shown as a red line. Accuracy distribution is shown as a swarm plot of 250 randomly subsampled reads. Average raw read quality of ONT reads is indicated by the color of the individual points.*

We also analyzed the same cDNA pool using a standard, heavily multiplexed ONT 1D2 run generating 5,904 full-length 1D and 1,142 1D2 cDNA reads, and the PacBio IsoSeq protocol generating 233,852 full-length cDNA Circular Consensus (CCS) reads. We aligned the resulting reads generated by each protocol to the SIRV transcript sequences using minimap2 and calculated percent identity (accuracy) using those alignments. The 1D2 run produced reads with a median accuracy of 87% (1D reads) or 95.6% (1D2 reads), while PacBio CCS reads had a median accuracy of 98.9%. R2C2 reads had a median accuracy of 94% (Fig. 4C) with the accuracy of individual R2C2 reads being highly correlated with average quality score of its underlying raw read as well as the numbers of subreads this raw read contained (Fig. 4C). While mismatch errors declined rapidly with increasing number of subreads, insertion and deletion errors declined more slowly. This might be explained by insertion and deletion errors not being entirely random but systematically appearing in stretches of the same base, i.e. homopolymers[35]. Indeed, 4-mers ('AAAA','CCCC','TTTT','GGGG') were enriched around insertion and deletion errors in R2C2 consensus reads. Overall, more aggressive filtering of R2C2 reads based on raw read quality score and subread coverage could increase the median accuracy of the R2C2 method but would also reduce overall read output.

Run Type	cDNA source	Raw Base output (Gb)	Raw Read output	Raw reads with length >1kb and Q ≥ 9	Full-length R2C2 Consensus reads
1D	SIRV E2	4.15	828,684	621,970	435,074
RAD4	B cells	2.06	408,347	227,250	149,791
RAD4	B cells	3.59	583,192	356,245	248,546
RAD4	B cells	4.23	877,412	528,800	345,402
RAD4	B cells	4.75	1,004,208	593,086	388,968

Table 1: R2C2 run statistics

R2C2 correlates well with PacBio for the quantification of SIRV transcripts

SIRV E2 transcripts vary in length from ~0.3-2.5 kb and are provided in four nominal concentration bins (“1/32”,”1/4",”1”,”4”) varying across two orders of magnitude. By analyzing the same SIRV E2 cDNA pools using R2C2 and PacBio IsoSeq we found that our R2C2 transcript counts generally matched nominal SIRV concentrations (Fig 5A). Additionally, there seems to be no clear length bias (Fig 5B), and our R2C2 transcript counts matched PacBio transcript counts very well with a Pearson correlation coefficient of 0.93 (Fig 5C).

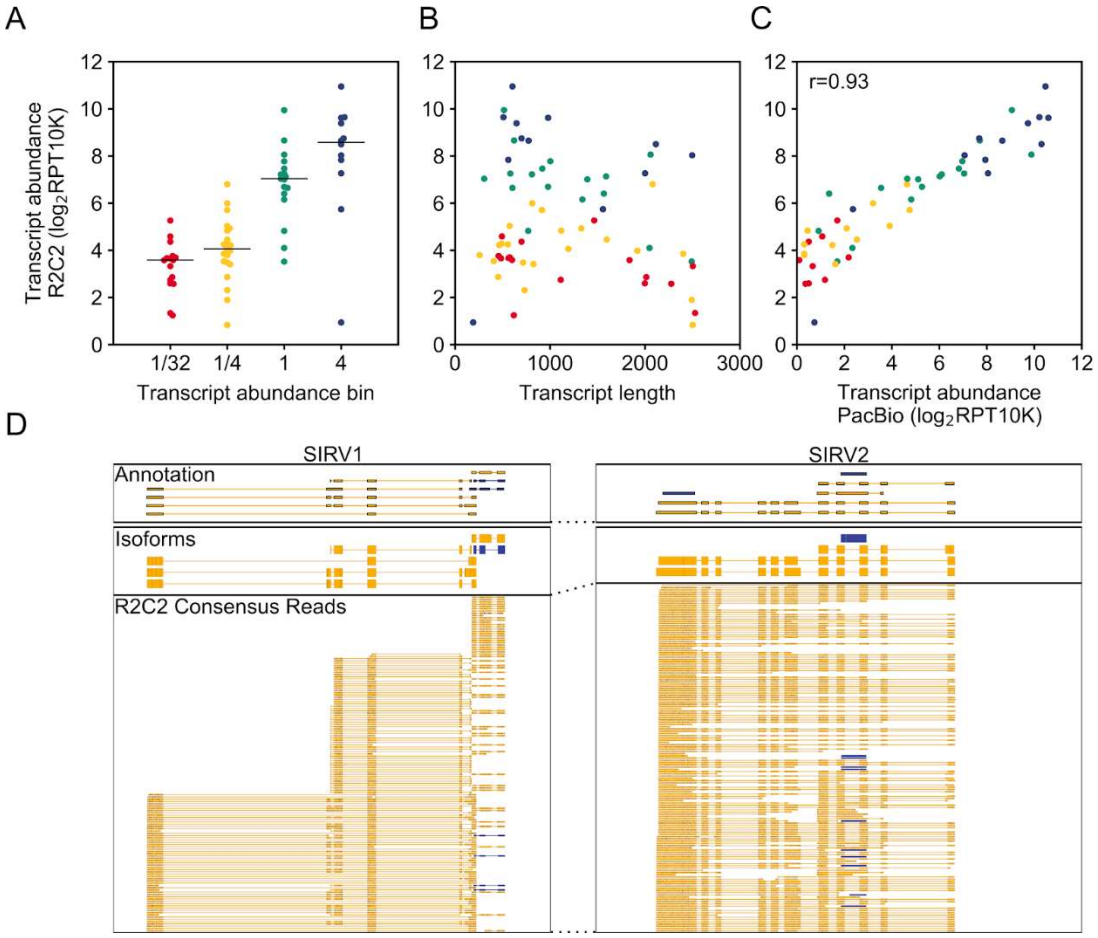

Fig. 5: R2C2 reads can quantify SIRV transcripts. *R2C2 reads were aligned to SIRV transcripts using minimap2 and expression values transcript abundance determined as Reads Per Transcript Per 10K reads (RPG10K). The transcript count ratio was plotted on the y-axis against A) the nominal transcript abundance bin reported by the SIRV transcript manufacturer (Lexogen), B) the transcript length, and C) transcript count ratio calculated from PacBio Isoseq reads. Pearson correlation coefficient (r) is reported in C). Each point represents a transcript and is colored according to it transcript abundance bin in all panels. D) Genome browser view of Transcriptome annotation, isoforms identified by Mandalorion, and R2C2 consensus reads is shown of the indicated synthetic SIRV gene loci. Transcript and read direction is shown by colors (Blue: + strand , Yellow: - strand)*

This indicates that the potential variation in transcript quantification seen in Figure 5A were either rooted in differences in the initial RNA concentration found in the SIRV E2 mix or

biases of our modified Smart-seq2 based cDNA amplification step rather than new biases introduced by the sequencing technology.

R2C2 enables simple and accurate isoform identification

Next we tested whether the increased accuracy of R2C2 reads would benefit splice junction and isoform identification. To this end, we aligned PacBio, ONT and R2C2 reads to the artificial SIRVome sequence provided as a genome reference for their SIRV transcripts (Fig. 4D). 91% of splice junctions in R2C2 consensus reads matched annotated splice sites perfectly, far exceeding ONT 1D raw reads at 80% and approaching PacBio CCS reads at 96%.

This increased accuracy allowed us to simplify our Mandalorion pipeline for isoform identification (see Methods). To test how this new version of Mandalorion would perform we subsampled R2C2 consensus read alignments to levels found in highly expressed genes in whole transcriptome analysis (500 read alignments per SIRV gene locus). Some of these subsampled R2C2 consensus reads did not align from end to end to a SIRV transcript (Fig. 4D). We suspect they are products of cDNA synthesis of degraded RNA molecules likely caused by repeated freeze-thaw cycles of the SIRV E2 standards for they all contained complete 5' and 3' priming sites and adapter sequences. This highlighted the importance of RNA integrity for full-length transcriptome sequencing. Indeed, R2C2 reads created from single B cell lysates which are thawed only once immediately before cDNA synthesis showed evidence of degradation products at much lower levels.

Because these degradation products appear to be largely random, they had little effect on the Mandalorion pipeline which identified 34 high confidence isoforms based on the subsampled R2C2 consensus reads (Fig. 4D). 24 of these isoforms matched annotated transcripts from the "1" and "4" abundance bins, while eight isoforms matched annotated transcripts from the "1/4" and "1/32" abundance bins. Only two high confidence isoforms represented truncated transcripts, caused by an oligodT mispriming on an A-rich region of the

SIRV303 transcript, or a premature template switch on the (likely degraded) SIRV602 transcript, respectively. This indicated that R2C2 consensus reads paired with the Mandalorion pipeline can identify complex transcript isoforms. It also highlights the difficulty of correct identification of low abundance transcript isoforms and the abiding problem of incomplete cDNA amplification.

R2C2 allows the demultiplexing of 7-8nt cellular indexes

Next we tested whether R2C2 reads are accurate enough to demultiplex reads based on short cellular indexes like those employed by 10X, Drop-Seq or our own Tn5Prime single cell RNAseq protocols. To this end, the SIRV cDNA we sequenced with the R2C2 method was indexed with 8 distinct combinations of a 7nt (TSO) and a 8nt (Nextera adapter) indexes. We found that we could confidently assign one 7nt and one 8nt index to 74% of R2C2 reads using a custom demultiplexing script based on Levenshtein distance between the observed sequence at the index position and our known input indexes. In 99.8% of the R2C2 assigned reads we found the combination of identified indexes matched one of the distinct combinations present in the cDNA pool.

Analysis of 96 single B cell transcriptomes using R2C2

Having established that we could demultiplex our Tn5Prime data using R2C2 reads with very little crosstalk between samples, we sequenced cDNA from 96 single B cells which we have recently analyzed using Illumina sequencing[40]. To streamline the sequencing reaction we used the ONT RAD4 (RAD004) kit which has a lower average read output than the ligation based 1D kit but has a much shorter (~20min) and, in our hands, more consistent and less error-prone workflow. Using the ONT RAD4 kit we generated 2,064,911 raw reads across 4 sequencing runs using R9.5 flowcells. C3POa generated 1,132,707 full-length R2C2 consensus reads which matched the length distribution of the sequenced cDNA closely (Figure 6A). 975,500 of the R2C2 consensus reads successfully aligned to the human

genome and 730,023 of those aligned reads were assigned to single B cells based on their 7nt and 8nt cellular indexes. We found that the vast majority of those reads were complete on the 5' end by comparing the alignment ends of these reads to transcription start sites (TSSs) previously identified[40] using Illumina sequencing. 653,410 of 730,023 (90%) reads either aligned to within 10bp of a predicted TSS (604,940 reads) or aligned within a rearranged antibody locus (48,470 reads) which makes accurate read alignment impossible.

R2C2 quantifies gene expression in single human B cells.

Individual cells were assigned 7,604 reads on average. We detected an average of 532 genes per cell (at least one R2C2 consensus read overlapping with the gene). Both the numbers of genes detected as well as gene expression quantification based on these R2C2 consensus reads closely matched RNAseq-based quantification[40]. When comparing gene expression of the same cell, RNAseq and R2C2 quantification had a median pearson correlation coefficient (r) of 0.79 opposed to 0.14 when comparing different cells with one another (Fig 6B). Using t-SNE clustering on R2C2 and Illumina data resulted in the sub-clustering of the same J chain-positive cells which we previously identified as plasmablasts (as opposed to memory B cells) (Fig. 6C).

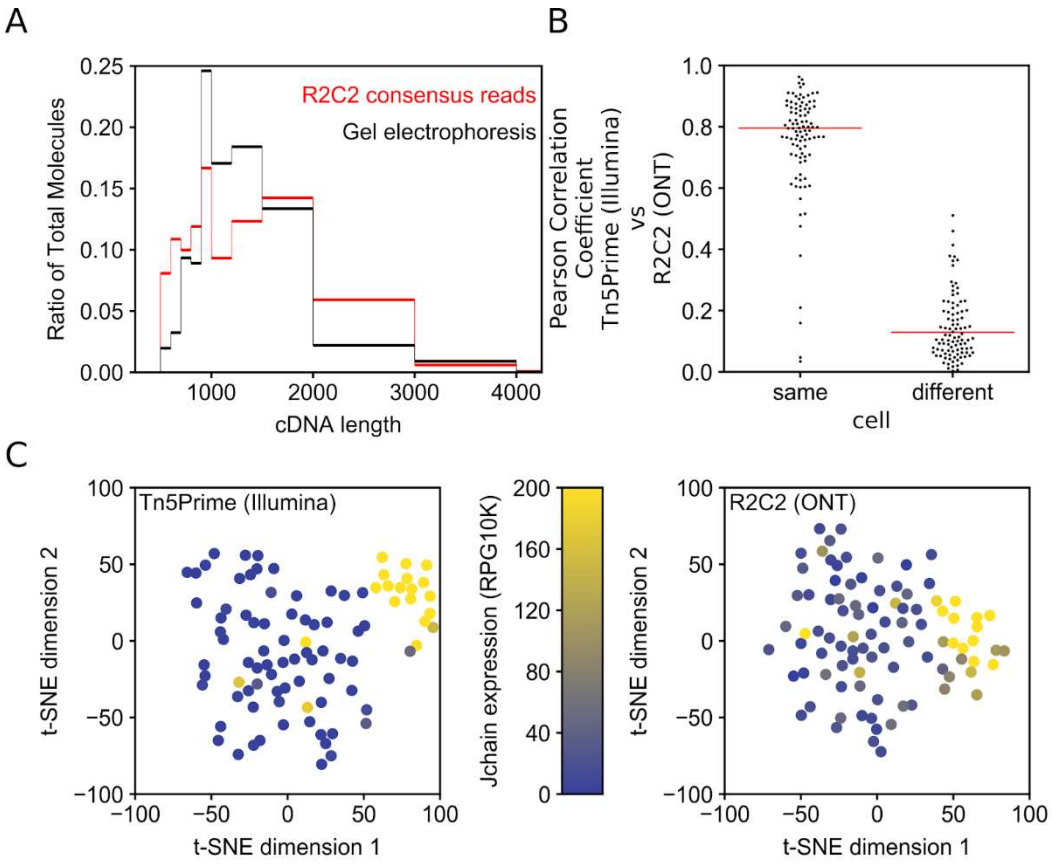

Fig. 6: R2C2 length bias and gene expression quantification. *A) B cell cDNA molecule length distribution as determined by electrophoresis on 2% agarose gel is compared to R2C2 consensus read length distribution. B) Pearson correlation coefficient (r) is shown for R2C2 and Illumina based gene expression quantification of the same of different cells. Red lines indicate medians. All 96 correlation coefficient from same cell comparisons and 96 subsampled correlation coefficients from different cell comparisons are shown as a swarmplot to display their distributions. C) t-SNE dimensional reduction plots of the same 96 B cells whose transcriptome were quantified with either the Tn5Prime Illumina based method or the R2C2 ONT based method. Cells are colored based on the Jchain expression which is strongly associated with plasmablast cell identity.*

R2C2 identifies isoforms in single human B cells.

We used our updated Mandalorion pipeline to identify high confidence isoforms separately for each of the 96 B cells we analyzed. By grouping R2C2 consensus reads based on their

splice sites and alignment starts and ends, Mandalorion identified an average of 163 high confidence isoforms per cell. We found that identification of high confidence isoforms was dependent upon R2C2 consensus read coverage. We identified at least one isoform in 3.1%, 64.9%, 92.2% of genes covered by 1-4 reads, 5-9 reads, or >10 reads, respectively. The vast majority of genes with >10 R2C2 consensus reads contained one (78%) or two (11%) isoforms, highlighting the low complexity of single cell transcriptomes.

Overall, the isoforms we identified had a 99.1% sequence similarity with the human genome. As previously observed for mouse B1 cells[26], human B cells show a diverse array of isoforms across their surface receptors. CD37 and CD79B, which were expressed in several B cells, showed diverse isoforms. These isoforms were defined by 1) intron retention events (CD79B: Cell A12_TSO6, CD37: Cells A11_TSO2 and A17_TSO1), 2) variable transcription start sites (TSSs), and alternatively spliced exons (CD79B: Cell A20_TSO2, CD37: Cell A17_TSO1), with the alternatively spliced exon being only partially annotated (Fig. 7).

Finally, for the B cell defining CD19 receptors we also observed multiple isoforms across cells, which is of particular interest because CD19 is a target for CAR T-cell therapy. Alternative splicing of CD19 has been shown to confer therapy resistance to B cell lymphomas. Interestingly, when we reference corrected (squanti-qc[44]) and translated the 4 isoforms we identified, only one contained the epitope required for FMC63 based CAR T-cell therapy (Fig. 7)[37–39].

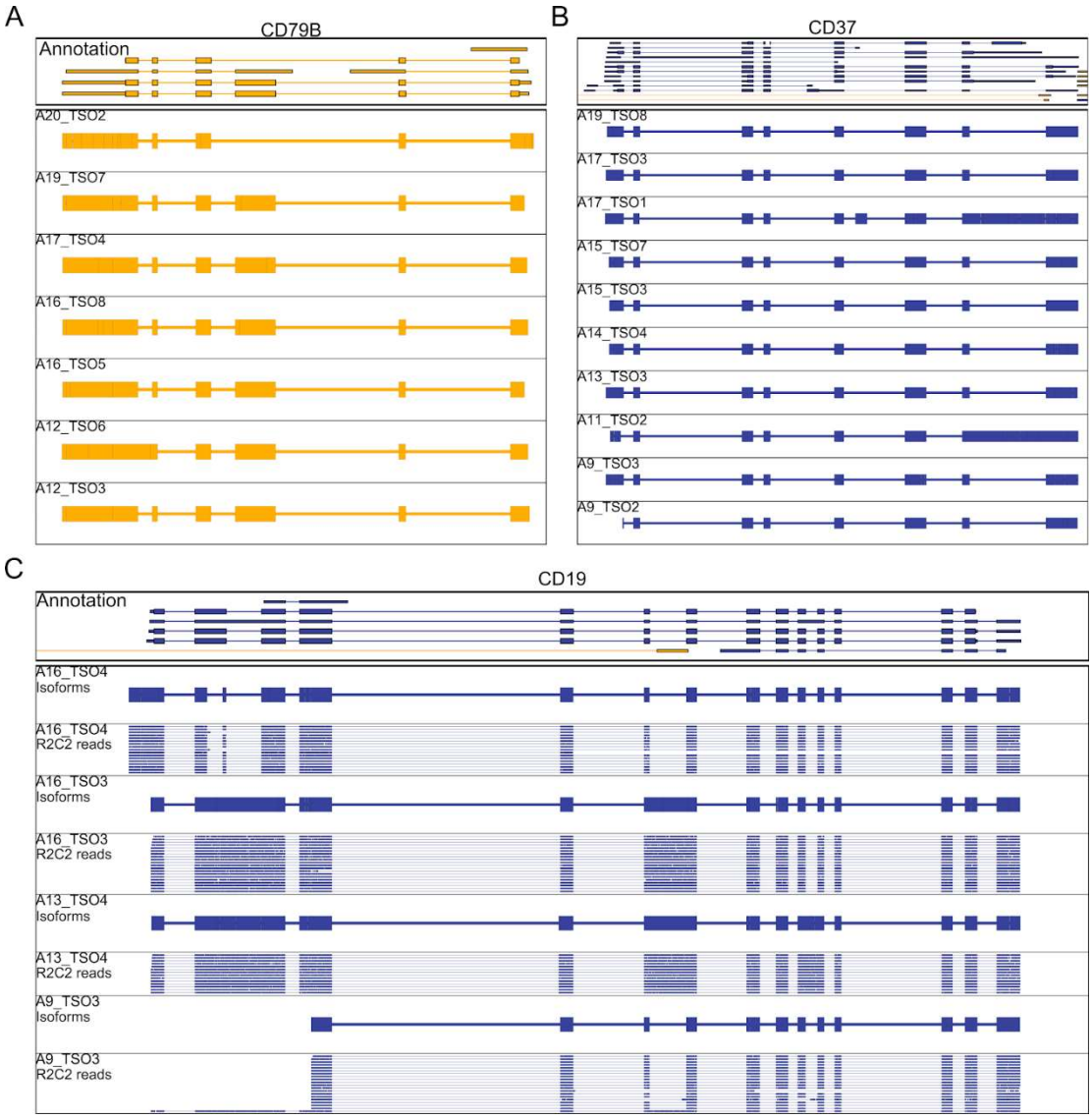

Fig. 7 R2C2 reads identify isoforms in B cell surface receptor genes

Genome browser views of Transcriptome annotation, isoforms identified by Mandalorion, and R2C2 consensus reads (C only, downsampled to 20 reads) are shown for the indicated gene loci. Transcript and read direction is shown by colors (Blue: + strand , Yellow: - strand). Cell IDs are indicated by combinations of A and TSO indexes.

Discussion

While RNAseq analysis has fundamentally changed how transcriptional profiling is performed, it is ultimately a stop-gap solution born from the limitations of short-read sequencing technologies. The need to fragment transcripts to fit short-read technologies like Illumina results in often unsurmountable analysis challenges. As a result, RNAseq analysis is often used like gene expression microarrays with the data used for downstream analysis being gene-expression values. Single cell RNAseq has further exacerbated this limitation because it is often restricted to 3' or 5' tag counting and generates gene expression values that are sparse due to both biological and technical reasons.

This results in a loss of information because individual genes can express many different isoforms, often with different functions. However, many bulk and single cell RNAseq methods do generate full-length cDNA as a intermediate product in library preparation. Long-read technology is able to take advantage of this full-length cDNA. While long-read sequencing technologies do not currently match Illumina's read output and accuracy, their outputs and accuracies are increasing. Here, we produced over 200,000 reads at close to 99% accuracy per run using the PacBio Sequel. Further, in our hands, the standard ONT $1D^2$ protocol can generate 1 million 1D cDNA reads at 87% accuracy and 50,000 $1D^2$ reads at 95% accuracy in a single run. The ONT based R2C2 sequencing method we developed takes advantage of this high throughput and increases ONT read accuracy. The R2C2 method we developed offers a compromise between PacBio and ONT technologies that generates on average 316,000 full-length cDNA reads at 94% accuracy in a single run. While the per run cost of flowcells and reagents of PacBio and ONT are roughly comparable, the capital cost of the PacBio Sequel sequencer (~$300k) vastly exceeds the cost of the ONT MinION (~$1k). This effective lack of capital costs associated with the ONT-based R2C2 method results in much lower total cost of accurate full-length transcriptome analysis compared to the PacBio IsoSeq workflow. Indeed, at its current throughput and accuracy and

combined with the low cost of the ONT MinION we believe that R2C2 brings comprehensive full-length transcriptome analysis within reach of most molecular biology laboratories.

In the immediate future, the R2C2 method will be a suitable complement for short-read sequencing. To this end, the R2C2 can be easily adapted to any RNAseq library preparation protocol that produces full-length double stranded cDNA molecules with known adapter/primer sequences at their ends. This includes standard Smart-seq2, 10X Genomics, and Drop-seq protocols. Adapting R2C2 to these protocols only requires the generation of a compatible DNA splint by modifying the primers used for amplifying the DNA splint. The same cDNA pool can then be sequenced by both Illumina and R2C2 methods.

We believe that R2C2 has the potential to replace short-read RNAseq and its shotgun approach to transcriptome analysis entirely, especially considering the impending wide release of the high-throughput ONT PromethION sequencer. This will be a significant advance considering the strength of full-length transcriptome sequencing showcased here. R2C2 paired with Mandalorion accurately identified full-length synthetic transcripts as well as several surface receptor isoforms of CD79B, CD37, and CD19 expressed by 96 distinct single human B cells. Identifying these full-length isoforms with short read RNAseq would have been impossible. Finally, the CD19 RNA isoforms we identified in the single B cells derived from a healthy adult may have implications regarding immunotherapy efficacy for most lacked the epitope in exon 4 that is targeted by FMC63 based CAR T-cell therapy. This confirms that even healthy individuals contain RNA isoform diversity for CD19 which may ultimately contribute to immunotherapy resistance when undergoing FMC63 based CAR T-cell therapy[37–39].

Methods

100pg of SIRV E0 (Lexogen) RNA or lysed single B cells (Collected from the blood of a fully consented healthy adult in a study approved by the Institutional Review Board (IRB) at UCSC) were amplified using the Tn5Prime[40] method, which represents a modification of the

Smart-seq2[11,45] method developed to capture 5' ends of transcripts using Illumina sequencing.

This method uses distinct template switch oligo (TSO) and oligodT primer sequences, enabling the easy differentiation of transcript 5' and 3' ends when using long-read sequencing. Following the Tn5Prime protocol, RNA or Single Cell Lysate were reverse transcribed (RT) using Smartscribe Reverse Transcriptase (Clontech) in a 10ul reaction including an oligodT primer and a Nextera A TSO containing a 7 nucleotide sample index (Table S1). RT was performed for 60 min at 42°C. The resulting cDNA was treated with 1 ul of 1:10 dilutions of RNAse A (Thermo) and Lambda Exonuclease (NEB) for 30min at 37°C. The treated cDNA was then amplified using KAPA Hifi Readymix 2x (KAPA) and incubated at 95°C for 3 mins, followed by 15 cycles for SIRV RNA or 27 cycles (single B cells) of (98°C for 20 s, 67°C for 15 s, 72°C for 4 mins), with a final extension at 72°C for 5 mins. cDNA amplification requires both the ISPCR primer and a Nextera A Index primer, which contains another 8 nucleotide sample index.

SIRV RNA: 8 SIRV E2 RNA aliquots were reverse transcribed and amplified in separate reactions adding distinct 7 nucleotide TSO and 8 nucleotide Nextera A Indexes to each resulting cDNA aliquot. The separate aliquots used directly as input into our R2C2 method or amplified using KAPA Hifi Readymix 2x (KAPA) (95°C for 3 mins, followed by 15 cycles (98°C for 20 s, 67°C for 15 s, 72°C for 4 mins), with a final extension at 72°C for 5 mins with ISPCR and Nextera_A_Universal Primers and pooled at equal amounts for input into PacBio Iso-Seq pipeline at the University of Georgia Athens sequencing core.

Single B cell lysates: Single B cells in separate in the wells of a 96 well plate were reverse transcribed using a distinct 7 nucleotide TSO index for each row. Columns were then pooled and amplified, using a distinct 8 nucleotide Nextera A Index for each pool. This resulted in the cDNA of all 96 cells carrying a unique combination of TSO and Nextera A index. This cDNA

was then pooled for Illumina sequencing (HiSeq4000 2x150)[40] or amplified using KAPA Hifi Readymix 2x (KAPA) (95°C for 3 mins, followed by 15 cycles (98°C for 20 s, 67°C for 15 s, 72°C for 4 mins), with a final extension at 72°C for 5 mins with ISPCR and Nextera_A_Universal Primers for input into our R2C2 method.

DNA splint amplification

A ~200bp DNA splint to enable Gibson Assembly[46] circularization of cDNA was amplified from Lambda DNA using KAPA Hifi Readymix 2x (KAPA) (95°C for 3 mins, followed by 25 cycles (98°C for 20 s, 67°C for 15 s, 72°C for 30 s) using primer Lambda_F_ISPCR(RC) and Lambda_R_NextA(RC) (Table S1). This generated a double stranded DNA with matching overlaps to full-length cDNA.

R2C2 sample preparation

Circularization of cDNA

200ng of cDNA was mixed with 200ng of DNA splint. Volume was adjusted to 6ul and 6ul of 2x NEBuilder Hifi DNA Assembly Master Mix (NEB). The reaction was incubated for 60min at 55°C. Volume was adjusted to 20ul and non-circularized DNA was digested using 1ul of 1:10 Exonuclease III and Lambda Exonuclease as well as 1ul of Exonuclease I (all NEB). Circularized DNA was extracted using SPRI beads with a size cutoff to eliminate DNA <500bp (0.8 beads:1 sample) and eluted in 50ul of ultrapure water.

Rolling circle amplification

Circularized DNA was split into 5 aliquots of 10ul and each aliquot was amplified in its own 50ul reaction containing Phi29 polymerase (NEB) and exonuclease resistant random hexamers (Thermo) (5ul of 10x Phi29 Buffer, 2.5ul of 10uM(each) dNTPs, 2.5ul random hexamers (10uM), 10ul of DNA, 29ul ultrapure water, 1ul of Phi29). Reaction were incubated

30°C overnight. All reaction were pooled and volume was adjusted to 300ul with ultrapure water. DNA was extracted using SPRI beads with a size cutoff to eliminate DNA <2000bp (0.5 beads:1 sample). At this point the High Molecular Weight DNA can easily shear. Therefore, beads and samples were mixed by gentle flicking of the tube, not vortexing or vigorous pipetting. Beads were allowed to settle for 5min on magnet, and after two 70% Ethanol washes, a mix of 90ul of ultrapure water, 10ul NEB buffer 2 and 5ul T7 Endonuclease was added to the beads. Beads were incubated for 2 hour on a thermal shaker at 37°C under constant agitation. Beads were then placed on magnet and supernatant is recovered. The DNA in the supernatant is then extracted again using SPRI beads with a size cutoff to eliminate DNA <2000bp (0.5 beads:1 sample) and eluted in 15ul of ultrapure water.

1ul of the eluate was diluted in 19ul of ultrapure water. 1ul of the 1:20 dilution was used to determine the concentration of the eluate using a Qubit High Sensitivity DNA kit (Thermo). The other 19ul were analyzed on a 1% agarose gel. Successful RCA and debranching by T7 Endonuclease results in HMW DNA that runs above the 10kb band of the NEB 2-log ladder but is not stuck in the loading well.

ONT sequencing

SIRV E2 RCA product was sequenced using the ONT 1D sample prep kit and a single 9.5 flowcell according to manufacturer's instructions with the exception that DNA was not sheared prior to library prep. Single B cell RCA product was sequencing using the ONT RAD4 kit and four 9.5 flowcells. The resulting raw data was basecalled using the albacore (version 2.1.3) read_fast5_basecaller script with the following settings:

1D run:

```
read_fast5_basecaller.py -r --flowcell FLO-MIN107 --kit SQK-LSK108 --output_format fastq --input /path/to/raw_data --save_path /path/to/output_folder --worker_threads 20
```

RAD4 runs:

```
read_fast5_basecaller.py -r --flowcell FLO-MIN107 --kit SQK-RAD004 --output_format fastq --input /path/to/raw_data --save_path /path/to/output_folder --worker_threads 20
```

C3POa data processing

Pre-processing (*C3POa_preprocessing.py*)

Basecalled raw reads underwent pre-processing to shorten read names and remove short (<1000kb) and low quality reads (Q<9) reads. Raw reads were first analyzed using BLAT[41] to detect DNA splint sequences. If one or more splint sequences were detected in a raw read, the raw read underwent consensus calling.

Consensus calling (*C3POa.py*)

1.) We identified tandem repeats in each raw read using a modified EMBOSS WATER[47–49] Smith Waterman self-to-self alignment. First, we set the ascending diagonal of the self-to-self alignment score matrix to 0, then we sum values across the all lines parallel to the diagonal. To speed up this self-to-self alignment, the score matrix is calculated in 1000 nucleotide bins. We then call peaks along these values which indicate the position of other splint sequences in the tandem repeats the raw read contains (Fig. 3B).

2.) Raw reads are then split into complete subreads containing full repeats and incomplete subreads containing partial repeats at the read ends. If there are more than 1 complete subreads, these complete subreads are aligned using poaV2[42] with the following command:

```
poa -read_fasta path/to/subreads.fasta -hb -pir path/to/alignments.pir -do_progressive NUC.4.4.mat >./poa_messages.txt 2>&1
```

The preliminary consensus is either reported by poaV2 (more than 2 subreads) or determined based on the poaV2 alignment by a custom script taking raw read quality scores into account (2 subreads). If only one complete subread is present in the raw read, its sequenced is used as consensus in the following steps.

3.) Complete and incomplete subreads are aligned to the consensus sequence using minimap2[50] and the following command

```
minimap2 --secondary=no -ax map-ont path/to/consensus.fasta
path/to/subreads.fastq > path/to/subread_overlap.sam
2> ./minimap2_messages.txt
```

4.) These alignments are used as input to the racon[43] algorithm which error-corrects the consensus.

```
racon --sam --bq 5 -t 1 path/to/subreads.fastq
path/to/subread_overlap.sam path/to/consensus.fasta
path/to/corrected_consensus.fasta > ./racon_messages.txt 2>&1
```

Post-processing (*C3POa_postprocessing.py*)

ISPCR and Nextera Sequences are identified by BLAT and the read is trimmed to their positions and reoriented to 5'->3'.

Alignment

Trimmed, full-length R2C2 reads and PacBio reads are aligned to the appropriate genomes and transcripts using minimap2. The following settings were used when:

Aligning to SIRV transcript sequences:

```
minimap2 --secondary=no -ax map-ont
/path/to/SIRV_Transcriptome_nopolyA.fasta
```

path/to/trimmed_corrected_consensus.fasta >

path/to/aligned.out.sirv.sam

Aligning to the “SIRVome” sequences:

minimap2 --splice-flank=no --secondary=no -ax splice

/path/to/SIRVome.fasta path/to/trimmed_corrected_consensus.fasta >

path/to/aligned.out.sirvome.sam

Aligning to the human genome (only chromosomes, no alternative assemblies, etc...):

minimap2 --secondary=no -ax splice

/path/to/hg38_chromosomes_only.fasta

path/to/trimmed_corrected_consensus.fasta >

path/to/aligned.out.hg38.sam

Percent identity of sequencing reads were calculated from minimap2 alignments. First md strings were added to the sam files generated by minimap using samtools calmd functionality. Matches, mismatches and indels are then calculated based on CIGAR and md string and percent identity is reported as *(matches/(matches+mismatches+indels))*100*.

For isofom identification and visualization SAM files were converted to PSL file format using the jvarkit sam2psl[51] script.

Isoform identification and quantification

Isoforms were identified and quantified using a new version of the Mandalorion pipeline (EII) with the following settings:

Isoform Identification:

python3 Mandalorion_define_and_quantify_isoforms.py -c

path/to/content_file -p path/to/output/ -u 5 -d 30 -s 200 -r 0.05 -R

3 -i 0 -t 0 -I 100 -T 60 -g /path/to/genome_annotation.gtf

Isoform alignment:

```
gmap -f psl -B 5 -t 6 -n 1 -d /path/to/human_reference_index
path/to/isoform_consensi.fasta > path/to/isoform_consensi.psl
```

Availability:

C3POa and Mandalorion will be available at github under https://github.com/rvolden/C3POa and https://github.com/rvolden/Mandalorion-Episode-II, respectively.

Raw read data are available at the SRA under PRJNA448331 (SIRV E2) and PRJNA415475 (B cells). Processed data are available at https://drive.google.com/file/d/1vP2EqJuXbN1TUIIXvUPZQmfSaUdKDgOr/view?usp=sharing

Aim 2: Improving C3POa performance

As outlined previously, the C3POa pipeline is divided into three discrete processing steps: pre-processing, consensus calling, and post-processing. The pre-processing step uses blat to align the DNA splint sequences to each read. The splint position information is used as a starting point (v0 and v1) for the read chunk in the self-to-self alignment. In v2, the splint alignment is only used for splint demultiplexing and determining the splint direction. The consensus calling step does the chunk-to-self/splint-to-read alignment to determine points of repetition in each raw read. The resulting sub-reads are aligned together using a partial order aligner to create a preliminary consensus sequence. The partial order alignments are also aligned using minimap2 to produce an overlap file for later use. The preliminary consensus sequence gets polished with Racon using the minimap2 overlaps to produce a highly accurate consensus sequence. The post-processing step aligns 3' and 5' adapters to the consensus sequences to ensure the output only contains full-length cDNA sequences.

While the original C3POa program (v0) was able to call consensus sequences given raw R2C2 reads, there were still many ways that it could be improved. Small optimizations were made over time from v0 to v1. One of the main differences was an updated aligner, gonk, which was written as a command line tool to do the chunk to splint alignment. The other main difference from v0 to v1 was the inclusion of built-in multiprocessing. While v1 was serviceable, it was clear that there were several technical disadvantages that needed to be addressed. The main technical disadvantages were an inability to handle internally repetitive sequences, a fragmented code base, and slow runtime.

The first disadvantage, C3POa's inability to handle internally repetitive sequences, is a result of the original alignment scheme that C3POa uses. Due to previous limitations in raw basecalled accuracy, we needed to align a large portion of the original read to itself. While this works well for cDNA, it does not translate well for genomic DNA. Genomic DNA has a tendency to be repetitive, which adds a lot of noise to our chunk-to-self alignment. Repetitive

genomic DNA is also why short reads are not well suited to resolve these problem regions. To get around this, it is more ideal to solely align the splint sequence to the raw read.

The second disadvantage was a fragmented code base. Previous versions of C3POa have three processing scripts: pre-processing, consensus calling, and post-processing. There was little cohesion between these three scripts, which led to user confusion and frustration. We also have auxiliary scripts that our lab would use internally to account for more complex library preparation methods (ie. adding hairpins to the sequencing prep and multiplexing samples with barcoded oligo-dT molecules). It would be more convenient to build these features into the existing scripts rather than having to run up to five different scripts to get usable data.

The last disadvantage is slow runtime. There are many sources of slowdown in all versions of C3POa. In v0, the main bottleneck was a lack of multiprocessing support. To get around this, we used GNU Parallel after splitting up the original input file. This was serviceable for a while, but was arduous to set up and it became clear that we needed built-in multiprocessing, which was introduced in v1. The most significant bottleneck was the amount of I/O required since there were various dependencies that were run through the command line within the python scripts. All of our command line dependencies (blat, water/gonk, poa, racon) were extremely slow to run because each time one of these was run, our script would need to write a file to input into the dependency and then read back in its input. In the best case (sequential access), writing to disk is about an order of magnitude slower than keeping data in memory. To get around the excessive waiting on I/O, it would be better to keep as much data as possible in memory.

Algorithmic Differences

C3POa v2 aims to alleviate all of the previously mentioned technical disadvantages. The first change made to the pipeline was to change how repetition is detected in raw reads. Instead of relying on aligning a 1kb chunk of the read to its entirety, we switched to only aligning our

200bp splint sequence, which was enabled due to better basecalling accuracy. Aligning a shorter sequence offers better specificity for terminal peaks, as seen in Figure 8. Another improvement that the v2 peaks provide is more accurate repetition positions. Because the alignment is only 200bp, the peaks are much narrower compared to the v1 peaks. Due to noise in the 1kb chunk alignment, the placement of each peak is less precise. The v1 chunk alignments were also susceptible to lumpy peaks as seen in Figure 3 and 4A. These lumpy peaks necessitated a custom peak caller, which has been replaced by the SciPy find_peaks module. This new peak finding method should solve the problem of not being able to reliably consensus call internally repetitive sequences, which should make R2C2 and C3POa better suited for genomic DNA sequencing.

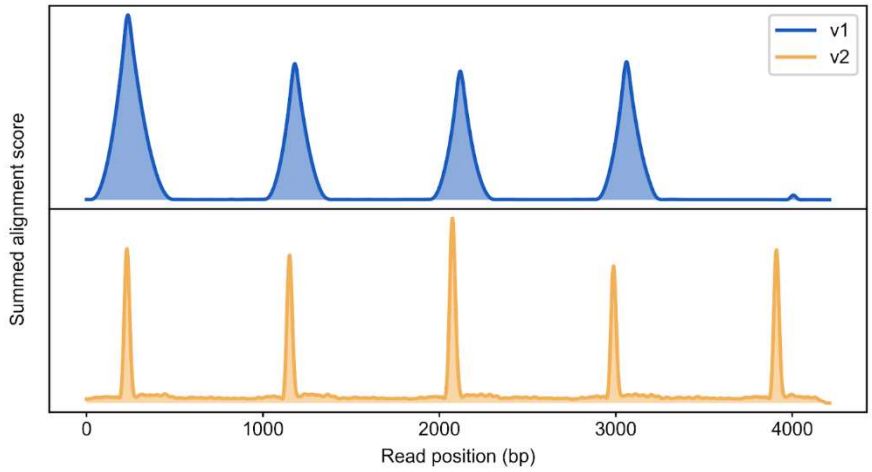

Fig. 8: C3POa peak calling update allows for more specificity. *Top panel shows peaks from C3POa v1, where the terminal peak at ~4000bp is not detected. Bottom panel shows peaks from C3POa v2, where the terminal peak is captured. The splint peaks in v2 allow for more accurate positioning because the peaks themselves are much narrower.*

The second pipeline change was to consolidate the various auxiliary scripts used to further process data as well as make some quality of life improvements. First off, the pre-processing portion is now part of the main consensus calling script. To accommodate this, we needed an easy way to skip the pre-processing step if it has been done previously. Because

the only information we need from the pre-processing is which splints align to which reads, we save the blat output. C3POa will also check for the blat output before doing the pre-processing so if it has been done before, it won't have to redo the alignments. For the post-processing, we ultimately decided to keep it as a separate script because of how much functionality was incorporated in v2. Instead of having separate scripts for orienting reads, demultiplexing oligo-dT barcodes, and detecting single-cell barcodes, all of it is handled by the post-processing script. As a quality of life improvement, we switched from our own custom FASTA/Q parsers to mappy's (the minimap2 Python interface) fastx_read API. This change allows us to easily and dependably read FASTA/Q files, even if they are compressed. To make compressed file handling more cohesive, we also added an option to write compressed output files. Adding compressed file compatibility has been immensely helpful because it allows us to practically half our storage by compressing all of the basecalled FASTQ files.

The last pipeline change was to improve the slow runtime, which in our case has three main facets: lack of multiprocessing, excessive disk I/O, and Python limitations. As previously stated, native multiprocessing support for the main consensus calling script was introduced in C3POa v1. However, both the pre-processing and post-processing were still single threaded. In C3POa v2, both have full multiprocessing support, which cuts down runtime dramatically. Next, we had to move away from using multiple command line tools within each script because of how slow it is to write to a file for an external program and then read its output. The first change was to rewrite and replace the old custom aligner, gonk. I wrote gonk to be a faster version of the originally modified Water from EMBOSS. While it was faster, wrapping Go code to be compatible with Python is less straightforward than using Cython. The new aligner, conk, is implemented in Cython and is imported as a library. Importing the aligner and keeping all of the scores in memory allows us to eliminate all disk I/O for our custom aligner. The next change was to the partial order aligner we were using, which was originally poaV2. We switched to abPOA, which on top of optimizing poa's

runtime, also has a Python interface (pyabpoa) that does not sacrifice functionality. We also switched from minimap2 to its Python interface, which allows us to generate subread overlaps without any disk I/O. Unfortunately, even using these Python interfaces does not completely eliminate disk I/O because Racon does not have any Python interface. This means we still need to write out the input files for Racon and read back in its input.

Results

After implementing all of the changes for C3POa v2, there were a couple of metrics that we wanted to benchmark. Most importantly, we wanted to see the runtime difference between C3POa v1 and v2. Considering the runtime of the entire pipeline, v1 takes approximately 8 hours where v2 takes 45 minutes (10.6x speed increase). We also wanted to look at how runtime is affected by the input sequence length, as seen in Figure 9. This is only measuring the number of seconds to consensus call a single read. Across the entire dataset, the median runtime for consensus calling using C3POa v1 was 1.77 seconds per read. For C3POa v2, the median runtime drops to 0.25 seconds per read (7x speed increase).

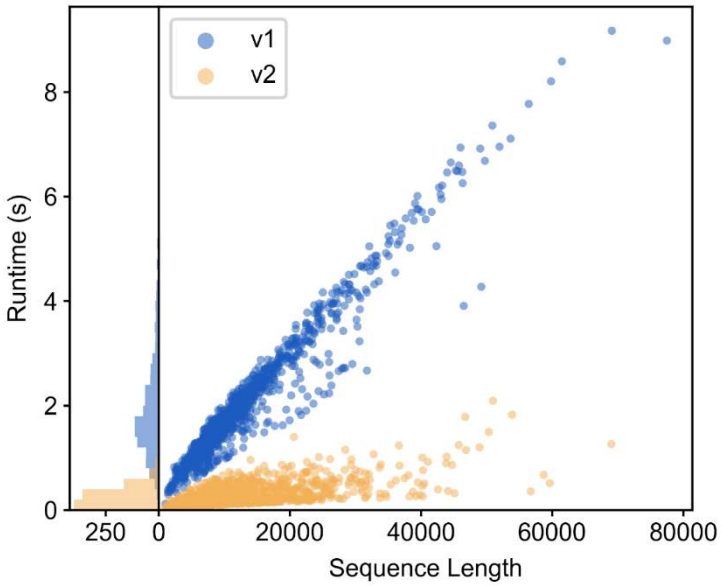

Fig. 9: Runtime difference between C3POa v1 and v2. *The median v1 runtime was 1.77 seconds per read, whereas the median v2 runtime was 0.25 seconds per read. The points plotted are a random subsample of 1000 reads for both C3POa versions.*

By reducing the amount of disk I/O and using improved dependencies like conk and pyabpoa, we were able to achieve an order of magnitude speedup across the whole pipeline.

Now that we have a more performant version of C3POa, we wanted to make sure the accuracy was at least on par with the previous versions. To evaluate base accuracy, reads are aligned to the human genome using minimap2. Using a custom script, it's possible to calculate the number of matched bases in the alignment over the length of the read. Here we present accuracy as a function of the subread coverage as calculated by C3POa in Figure 10. It is important to note that the calculated coverage is dependent on the number of complete subreads found during the splint to chunk alignment.

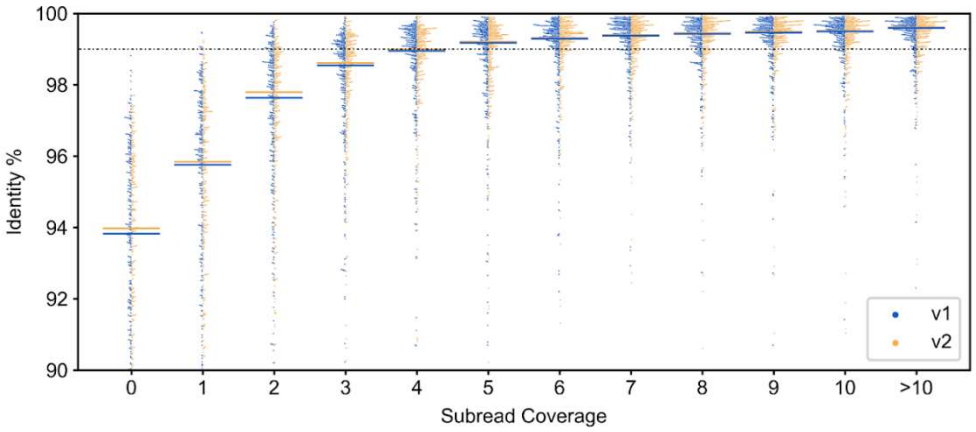

Fig. 10: Base accuracy for each coverage bin for C3POa v1 and v2. *C3POa v2 has very slightly higher accuracy on the median than v1. Impressively, we get Q20 with only 4x coverage on a single molecule.*

Something curious about the accuracy is that we see a small increase in bins 0 through 2. For these bins, the accuracy bump is most likely due to the switch from poa to abPOA as well as a fixed pairwise consensus calling script. Another factor that could cause a subtle bump in accuracy is the specificity of the splint to read alignment. Because only full length subreads are included in the preliminary consensus, reads with terminal peaks will end up being less accurate using C3POa v1. Compared to the previous version of C3POa, v2 also increases throughput. In the case of single-cell cDNA sequencing with 10X Genomics, the throughput of post-processed reads increased by 20% from C3POa v1 to v2. Using this updated C3POa version, we were able to consensus call more reads to higher accuracy faster.

Aim 3: Single-Cell cDNA Isoform Sequencing with 10X Genomics and R2C2

This section is adapted from **Highly Multiplexed Single-Cell Full-Length cDNA Sequencing of human immune cells with 10X Genomics and R2C2** (Volden et al. 2021).

Highly Multiplexed Single-Cell Full-Length cDNA Sequencing of human immune cells with 10X Genomics and R2C2

Roger Volden[1] and Christopher Vollmers[1,#]
1. Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064
[#]) Correspondence should be addressed to: Dr. Christopher Vollmers: vollmers@ucsc.edu

Abstract

Single cell transcriptome analysis elucidates facets of cell biology that have been previously out of reach. However, the high-throughput analysis of thousands of single cell transcriptomes has been limited by sample preparation and sequencing technology. High-throughput single cell analysis today is facilitated by protocols like the 10X Genomics platform or Drop-Seq which generate cDNA pools in which the origin of a transcript is encoded at its 5' or 3' end. These cDNA pools are currently analyzed by short read Illumina sequencing which can identify the cellular origin of a transcript and what gene it was transcribed from. However, these methods fail to retrieve isoform information. In principle, cDNA pools prepared using these approaches can be analyzed with Pacific Biosciences and Oxford Nanopore long-read sequencers to retrieve isoform information but all current implementations rely heavily on Illumina short-reads for the analysis in addition to long reads. Here, we used R2C2 to sequence and demultiplex 9 million full-length cDNA molecules generated by the 10X Chromium platform from ~3000 peripheral blood mononuclear cells (PBMCs). We used these reads to – independent from Illumina data – cluster cells into B

cells, T cells, and Monocytes and generate isoform-level transcriptomes for these cell-types. We also generated isoform-level transcriptomes for all single cells and used this information to identify a wide range of isoform diversity between genes. Finally, we also designed a computational workflow to extract paired adaptive immune receptor – T cell receptor and B cell receptor (TCR and BCR) – sequences unique to each T and B cell. This work represents a new, simple, and powerful approach that – using a single sequencing method – can extract an unprecedented amount of information from thousands of single cells.

Introduction

The analysis of transcriptomes using high-throughput sequencers has revolutionized biomedical research[6,52]. Pairing transcriptome analysis with the high-throughput processing of single cells has provided unprecedented insight into cellular heterogeneity[53,54]. Among many other studies, researchers have leveraged the strengths of high-throughput single-cell transcriptome analysis to create single cell maps of the mouse[30,55] or C. elegans[56] model organisms, to elucidate a new cell type in the lung involved in cystic fibrosis[57], and to increase our knowledge of adaptive and innate immune cells[58–61].

High-throughput single-cell transcriptome analysis however comes with trade-offs. In particular, droplet- or microwell-based methods like Drop-seq[62], 10X Genomics[63], and Microwell-Seq[55] or Seq-Well[64] single cell workflows generate pools of full-length cDNA with either the 5’ or 3’ end containing cellular identifiers. The cDNA pools are intended for high-throughput short-read sequencing and must therefore be fragmented such that one read sequence includes the cellular identifier and the sequence of its pair includes a fragment from within the original cDNA molecule. As a result, only a relatively short fragment of the cDNA is then sequenced alongside the cellular identifier limiting the resolution of this approach to the identification of genes associated with a given molecular identifier.

Instead of sequencing transcript fragments, long-read sequencing methods in the form of Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are now capable of sequencing comprehensive full-length transcriptomes[20–22,65]. These methods have now been used to analyze single cell cDNA pools generated by different methods, both well-[25,26,66] and droplet-based[24,67,68], enriching the information we can extract from single cells experiments. However, for the analysis of high-throughput droplet-based experiments with long reads, short-read data are still required for interpreting experimental data[24] or enabling the identification of cellular and molecular identifiers in low-accuracy ONT reads[67]. Short-read data remain a requirement because either long-read data are not of sufficient depth to cluster cells into cell-types or not accurate enough to decode cellular origin of cDNA molecules.

Because decoding the cellular origin of a cDNA molecule requires accurate sequencing of the molecular identifier, error-prone long read technologies are generally not sufficient to sequence each cDNA pool and to accurately interpret the single-cell data encoded therein. We have recently developed and applied the R2C2 approach which uses concatemeric consensus sequencing to improve ONT read accuracy from ~92% to 98% while still producing more than 2 million full-length cDNA sequences per MinION flow cell[25,65,69,70]. The combination of these technologies therefore has the potential to illuminate isoform-level single cell biology with unprecedented resolution.

In this manuscript we demonstrate that this combination of high throughput and accuracy is sufficient for the Illumina short-read independent analysis of highly multiplexed 10X Genomics cDNA pools. To this end we independently analyzed two pools containing the cDNA molecules of ~1500 human Peripheral Blood Mononuclear Cells (PBMCs) with Illumina and R2C2 (ONT) workflows. We showed that the R2C2 approach identifies the same cellular identifiers in the cDNA pools and generates comparable single-cell gene expression profiles and cell-type clusters. In addition, and in contrast to Illumina data, R2C2 data also allow the determination of cell-type specific and single-cell isoform-level transcriptomes. Finally, R2C2

allowed us to resolve and pair full-length adaptive immune receptors (AIR) transcripts in the B and T cell subpopulations of our PBMC sample which currently requires specialized library preparation methods and sequencing approaches.

Results

We extracted PBMCs from whole blood and processed the cells in replicate using the Chromium Single Cell 3' Gene Expression Solution (10X Genomics) aiming to include 1500 cells each for two replicates. We then divided the full-length cDNA intermediate generated by the standard 10X Genomics protocol to perform both short- and long-read sequencing (Figure 11A).

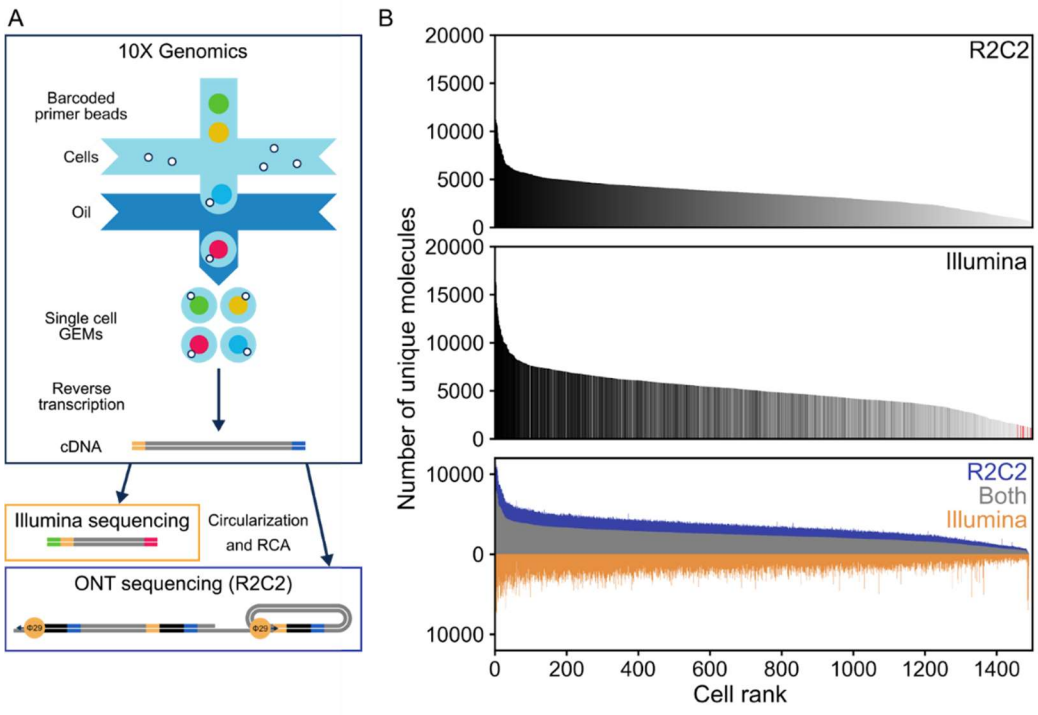

Fig. 11: Data Generation and Characteristics. *A) Thousands of peripheral blood mononuclear cells (PBMCs) were processed using the 10X Genomics Chromium Single Cell 3' Gene Expression Solution. The resulting full-length cDNA was either fragmented for Illumina sequencing or processed using the R2C2 workflow. B) After read processing and demultiplexing, the unique molecular identifiers (UMIs) associated with each cellular index (cell) in R2C2 (top) and Illumina (center) datasets are shown as histograms. Cells are ranked by the number of UMIs and colored based on their rank in the R2C2 dataset. Red lines indicate cellular identifiers found in Illumina but not R2C2 data. At the bottom, the UMIs shared between cellular identifiers in Illumina and R2C2 datasets or unique to each dataset are shown as stacked histograms. Cells are ranked by the number of shared UMIs. Data for replicate 1 are shown.*

Illumina data covers 10X UMIs comprehensively

For sequencing on the Illumina NextSeq, we fragmented the full-length cDNA according to the standard 10X protocol. We demultiplexed and merged the resulting reads based on cellular barcodes and unique molecular identifiers (10X-UMIs) associated with every amplified transcript molecule during reverse transcription (see Methods). By only keeping transcript molecules with a raw read coverage of >3, we condensed 202,469,707 raw read pairs to 15,264,862 reads originating from the 3’ ends of unique transcript molecules across both replicates (~5000 molecules per cell).

R2C2 data identifies the same cellular and molecular identifiers as Illumina data

For sequencing on the ONT MinION and PromethION sequencers, we processed 10ng of full-length cDNA using the previously published R2C2 workflow (see Methods). The resulting R2C2 libraries were then sequenced using standard ONT LSK-109 ligation-based sequencing kits. We processed the resulting ONT raw reads into R2C2 consensus reads using the C3POa pipeline (Table 2). We then merged reads in two sequential steps if they contained matching unique molecular identifiers (UMIs) in either the dsDNA splint used to

circularize cDNA molecules (Splint-UMI) or the 10X oligo(dT) primer used to prime reverse transcription of poly(A) RNA molecules (10X-UMI).

First, we merged 3.3% and 6.5% of the R2C2 consensus reads in replicate 1 and replicate 2 respectively because their Splint-UMI identified them as originating from the circularization of the same cDNA molecule. Second, we merged 46.3% and 46.1% of these Splint-UMI merged R2C2 consensus reads in replicate 1 and replicate 2, respectively, because their 10X-UMI identified them as originating from the same RNA molecule. Across both replicates this sequential merging process resulted in 14,822,072 Splint/10X-UMI merged R2C2 consensus reads with a median sequence accuracy of 98.0%.

Next, we demultiplexed these ~14.8 million Splint/10X-UMI merged R2C2 consensus reads based on the 10X cellular barcodes they contained. In this way, 81% of these reads could be successfully assigned to an individual cell, which compares favorably to the ~6% Illumina-independent and ~67% Illumina-guided assignment rates determined for standard ONT reads in previous studies[67,71].

	Basecalled reads	**Post-processed R2C2 consensus reads**	**Splint-UMI merged R2C2 consensus reads**	**Splint/10X-UMI merged R2C2 consensus reads**	**Demultiplexed R2C2 reads**
Replicate 1	29,529,179	11,564,494 (39.2%)	11,368,091 (98.3%)	7,853,440 (69.1%)	6,385,901 (81.3%)
Replicate 2	26,526,607	10,661,139 (40.2%)	10,276,420 (96.4%)	6,968,632 (67.8%)	5,652,620 (81.1%)

Table 2: Read numbers throughout processing.

Moreover, 2974 (99.1%) of the 3000 cellular identifiers we determined independently from the R2C2 dataset also appeared in the Illumina dataset.

Because we merged reads in Illumina and R2C2 datasets based on the 10X-UMI, each read in either dataset should originate from a unique RNA molecule. Consequently, the

number of reads assigned to each cell was also highly similar between the datasets (Fig 11B). Also, for each cell, 67% of the R2C2 reads contained a 10X-UMI that was also present in an Illumina read assigned to the same cell. Interestingly, the accuracy of R2C2 reads containing 10X-UMIs present in an Illumina read was significantly higher than the accuracy of R2C2 reads containing 10X-UMIs not present in an Illumina read (98.4% vs. 97.1%; p=0.0 Monte-Carlo Permutation test). This indicates that read accuracy plays an important role in accurately identifying UMI sequences. However, although their RNA molecule of origin cannot be unambiguously identified, R2C2 reads containing UMIs with sequencing errors are still highly valuable for downstream analysis.

Clustering single cells into cell types based on gene expression

We next investigated whether these R2C2 reads could be used to determine gene-expression accurately enough to cluster single cells into cell types – an analysis step that is currently routinely performed using short-read based gene expression. To this end, we used minimap2 to align R2C2 reads to the human genome (hg38) and used featureCounts to determine gene expression levels in each cell[72,73]. For comparison, Illumina reads generated from the same cDNA were aligned using STAR and also processed using featureCounts[74]. Median Pearson-r values for R2C2 and Illumina-based gene expression for the same cell showed high correlation at 0.74.

We then clustered R2C2 and Illumina datasets independently using the Seurat analysis package[75]. R2C2 and Illumina datasets both grouped into three cell type clusters. Based on marker gene expression, the major cell types could be identified as B cells (CD79A)[76], T cells (CD7)[77], and Monocytes (IL1B)[78] – the expected composition of a PBMC sample (Fig. 13). Importantly 99.5% of cells that were clustered in both datasets associated with the same cell type in the two datasets.

This showed that R2C2 reads show performance comparable to Illumina data for determining gene expression and clustering cell types in massively multiplexed single-cell experiments.

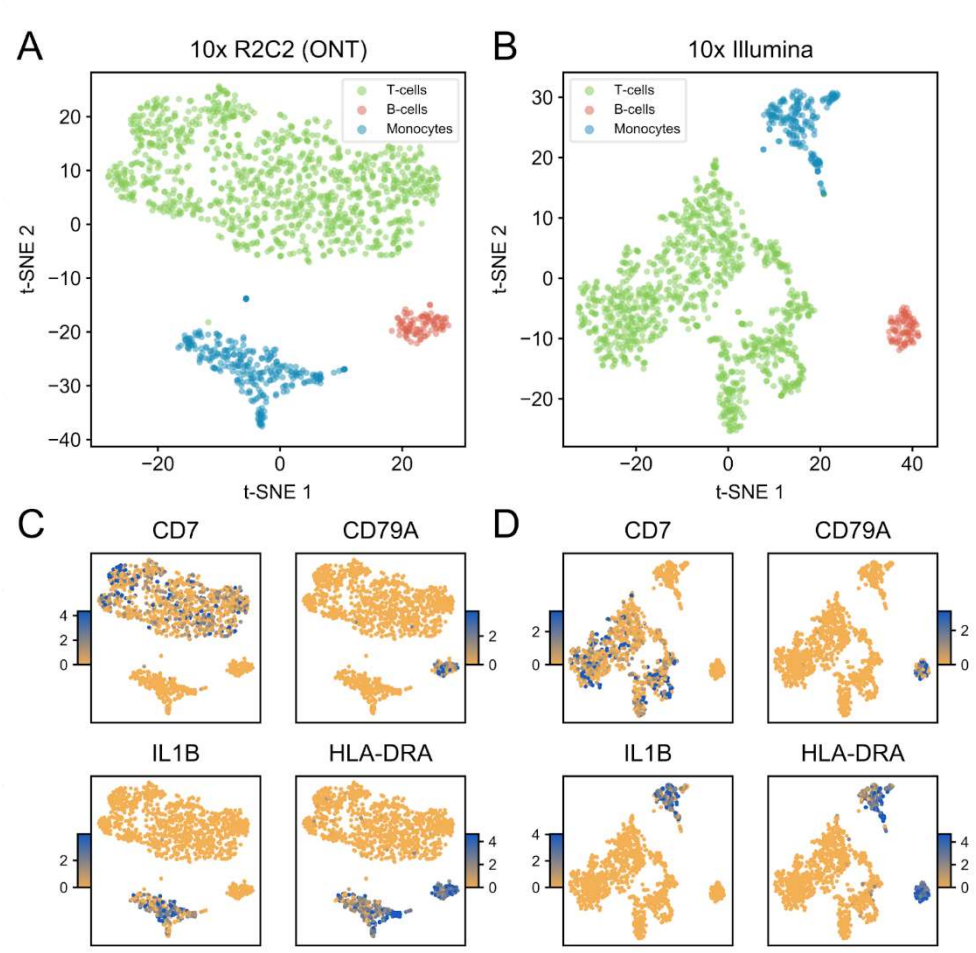

Fig. 12: R2C2 and Illumina datasets independently cluster into B cells, T cells, and Monocytes. *Gene expression profiles were determined independently for each cell in R2C2 and Illumina datasets. The Seurat package was then used to cluster cells based on the gene expression profiles. The cells in R2C2 (A) and Illumina (B) datasets both clustered into 3 groups which, based on marker gene expression (C and D) could be identified as B cells, T cells, and Monocytes. The color gradient (C and D) encodes ln(fold change), where the fold change is comparing that cluster's expression to the rest of the data. Data for replicate 1 are shown.*

Generating cell type specific isoform-level transcriptomes

Having successfully sorted cells into cell types, we set out to generate high quality transcriptomes for these cell types. This is possible because, as shown in previous studies analyzing 10X cDNA with long reads[24,67], R2C2 reads appeared to cover entire transcripts.

First, as previously established[24], we pooled all reads associated with the cells of each cell type to create a synthetic bulk sample. We then identified transcript isoforms for each synthetic bulk cell type using Mandalorion[25,26,65,69]. The majority (50-60%) of isoforms generated by Mandalorion for the individual cell types were classified by SQANTI[79] as either 'full-splice-match' or 'novel-in-catalog' which represent likely full-length isoforms. This number increased to >80% if only multi-exon isoforms were considered. In aggregate, the cell type specific isoforms we generated represent full-length B cell, Monocyte, and T cell transcriptomes, with each transcriptome's depths dependent on the number of cells and reads associated with each cell type (Table 3). With ~9 million R2C2 reads and 14,925 multi-exon isoforms, the T cell transcriptome is the most complete and likely most useful of the three cell types.

Cell type	Number of cells	Number of reads	Number of genes with multi-exon isoforms	Number of multi-exon isoforms
B cells	179	625,334	1,481 (plus 55 novel genes)	2006
T cells	2,199	9,108,828	6,934 (plus 448 novel genes)	14,925
Monocytes	464	2,042,162	2,882 (plus 77 novel genes)	4530

Table 3: Cell type specific full-length transcriptome characteristics

Differential isoform usage between cell types

In addition to determining which isoforms are expressed, we can also quantify the expression of these isoforms and investigate whether they are differentially expressed between the three cell types. To perform this differential isoform expression analysis, we first wanted to capture

all the isoforms expressed in the entire dataset. To this end, we composed an additional synthetic bulk sample using the R2C2 reads from all cells in the dataset. We then used Mandalorion to identify all isoforms present in this synthetic bulk sample and quantified the expression of each isoform in B cells, T cells, and Macrophages. Next, quantified isoforms were grouped by the genes they were associated with and genes with significant isoform usage between cell types were determined using a Chi-square contingency table test. After filtering for genes expressed in at least two cell types and multiple testing correction, we identified 74 genes with differential isoform usage (p-value<0.01). The features that distinguished differentially expressed isoforms included alternative TSSs (AIF1, Fig. 13B), cassette exons (CD83, Fig. 13C), or poly(A) sites (EIF4A1, Fig. 13D).

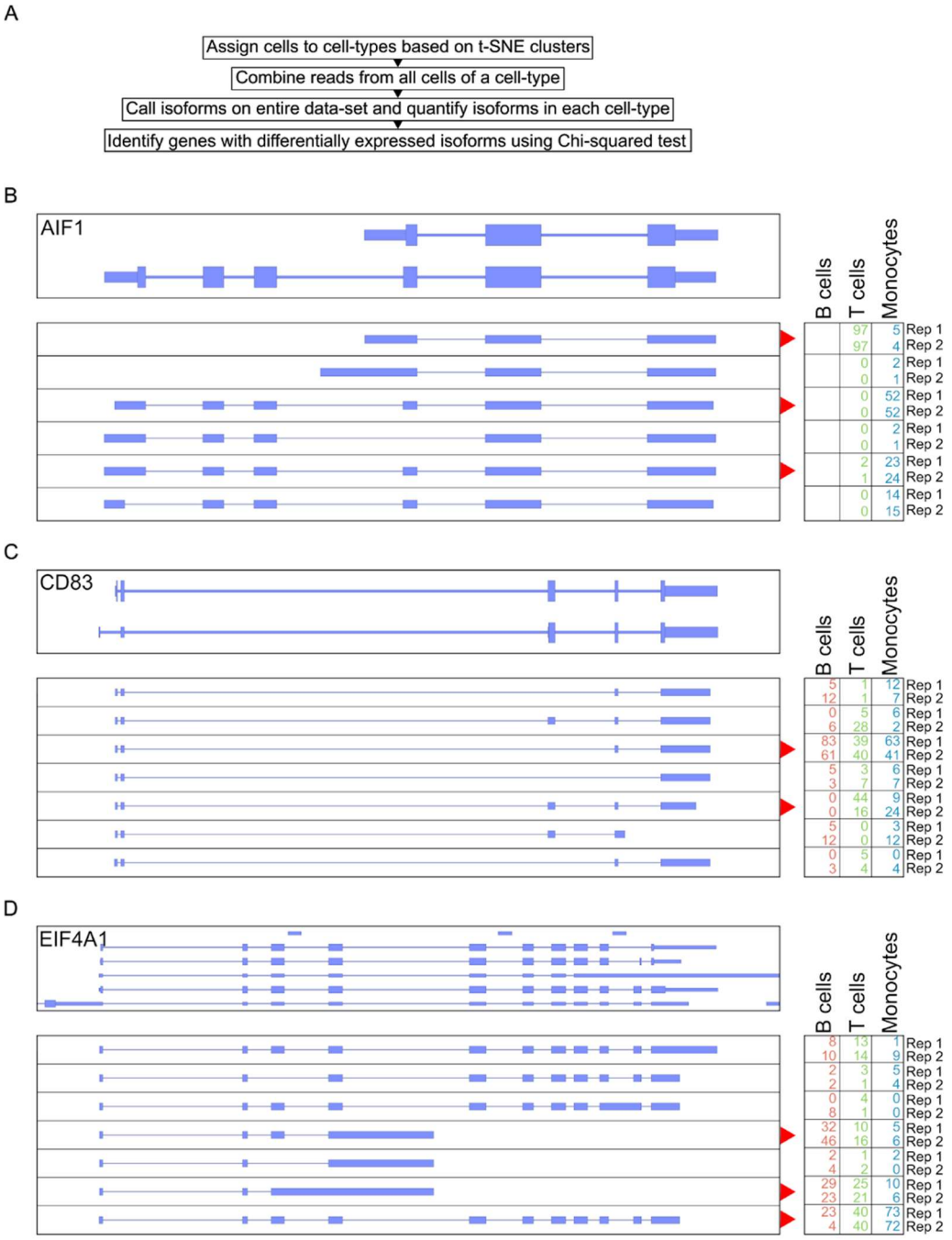

Fig. 13: Identifying differentially expressed isoforms between celltypes using clustered single cell data.

A) Workflow of differentially expressed isoform identification. R2C2 reads are separated by cell type, then used to identify and quantify isoforms. Genes with differential isoform usage between cell types are then identified using Chi-squared tests. B-D) Genome Browser shots of three genes with differential isoform expression. Gene annotation is shown on top. Isoforms as determined by Mandalorion on the entire dataset are shown below ("top strand"=blue, "bottom strand"=yellow). Relative quantification (%) of each isoform in each cell type and replicate is shown on the right. Isoforms with the most variable changes in abundance are indicated with a red arrow.

Isoform diversity is highly variable between genes

Next, we investigated whether single-cell derived transcriptome information can enrich our understanding of isoform diversity. While pooling all reads associated with a cell type can serve as a basis for defining transcriptome annotations, this approach loses information on which isoforms are expressed by which individual cell and due to coverage cut-offs likely presents a conservative estimate of the true isoform diversity present in a cell type.

In the 3000 cell dataset we present here, we have sufficient coverage to generate isoforms for each cell independently. Using Mandalorion, we generated a median of 127 multi-exon isoforms per cell, with the majority being classified as either 'full-splice-match' (77%) or 'novel-in-catalog' (11%).

We then analyzed isoform diversity across ~3000 the cells in the dataset. To this end, we merged identical isoforms expressed by different cells. We then determined how many cells expressed isoforms for any given gene.

Interestingly, isoform diversity varied greatly between genes (Fig. 14A). On one end of the spectrum, genes encoding ribosomal proteins in particular are expressed in the majority of cells, yet we identify few unique isoforms for these genes. For example, 1299 cells expressed a total of 1299 isoforms (as determined by Mandalorion) of the ribosomal protein gene RPL35. After merging all identical isoforms, only 8 unique isoforms remained and only

one of those was expressed by more than one cell. On the other end of the spectrum, genes like LMNA are also expressed by a majority of cells but feature many unique isoforms. In fact, 930 cells expressed a total of 969 unique LMNA isoforms. After merging all identical isoforms, only 305 unique isoforms remained and 86 of those were expressed by more than one cell.

Unique isoforms expressed by more than one cell as determined by this *'merged single cell'* approach could therefore be used to enrich isoform annotations based on bulk or synthetic bulk data. For example, combining all R2C2 reads collected for all the cells in this study and identifying isoforms based on this synthetic bulk yielded one isoform for RPL35 but also only 3 isoforms for LMNA, likely due to minimum relatives abundance requirements of 1% at a locus set as default in Mandalorion. In fact, most genes expressed by many cells had a low number of isoforms identified by the '*synthetic bulk'* approach (Fig. 14B).

By systematically comparing the number of isoforms determined by *'merged single cell'* and '*synthetic bulk*' approaches we showed that the more cells expressed isoforms for a gene, the more likely the *'merged single cell'* approach was to identify additional isoforms. This analysis highlighted the behavior of HLA class I genes, in particular HLA-B, HLA-C, and HLA-E (Fig. 14C), which all showed >40 isoforms with the *'merged single cell'* approach but only one or two in the *'synthetic bulk'* approach (Fig. 14A, B, D).

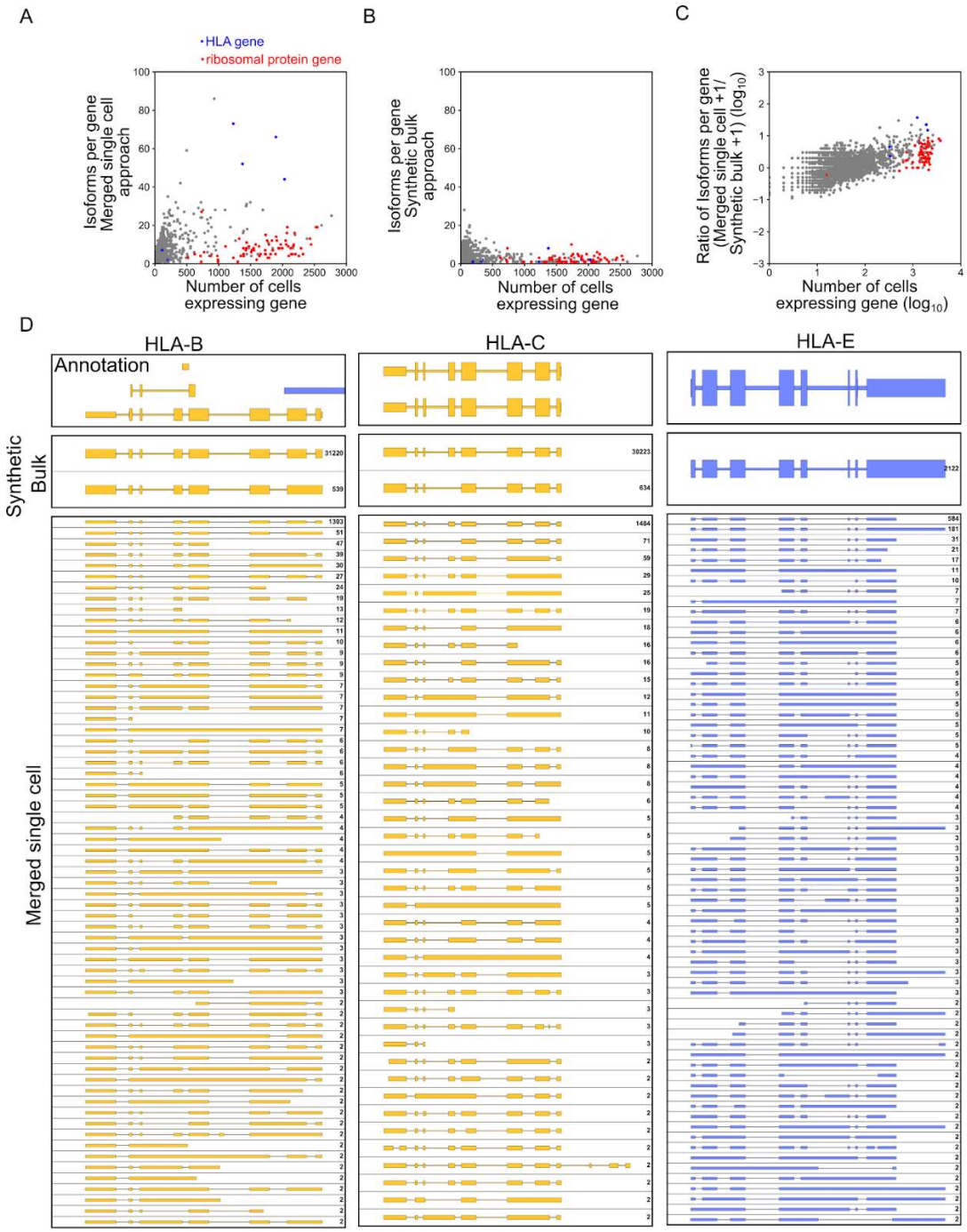
A
HLA gene
ribosomal protein gene
Isoforms per gene
Merged single cell
approach
Number of cells
expressing gene
B
Isoforms per gene
Synthetic bulk
approach
Number of cells
expressing gene
C
Ratio of Isoforms per gene
(Merged single cell +1/
Synthetic bulk +1) (log10)
Number of cells
expressing gene (log10)
D
HLA-B
HLA-C
HLA-E
Annotation
Synthetic
Bulk
Merged single cell

Fig. 14: Genes show a wide range of isoform diversity. *We generated an isoform level transcriptome for each cell in our dataset and then analyzed the isoform diversity for different genes by merging these isoforms. A) The correlation of the number of cells expressing an isoform for a gene and how many unique isoforms we identified for that gene using the 'merged single cell' approach is shown as a scatter plot. B) The correlation of the number of cells expressing an isoform for a gene and how many unique isoforms we identified for that gene using the synthetic bulk approach is shown as a scatterplot. C) The correlation of the number of cells expressing an isoform for a gene the ratio of the number of isoforms identified for that gene with the 'merged single cell' and 'synthetic bulk' approaches. Both number of cells and isoform ratio are shown as* log_{10}*. ABC) Genes encoding ribosomal proteins and HLA proteins are shown in red and blue respectively. D) Genome Browser shots HLA genes are shown. Genome annotation is shown on top, isoforms determined by the synthetic bulk approach in the middle, and isoforms determined by the merged single cell approach at the bottom. ("top strand"=blue, "bottom strand"=yellow). Number of reads (synthetic bulk) or cells (merged single cells) associated with an isoform are shown on the right.*

Extracting paired adaptive immune receptor sequences from B and T cells

In addition to the analysis of regular transcript isoforms, we investigated whether our datasets enable the identification and pairing of adaptive immune receptor (AIR) transcripts. AIR transcripts encode for antibodies and T cell receptors which pose unique challenges for sequencing applications. Each antibody (IG) or T cell receptor (TR) is encoded by two AIR transcripts each of which is transcribed from a gene whose V (, D,) and J segments are uniquely rearranged in each individual B or T cell.

Our standard Mandalorion transcript isoform identification workflow does not capture these AIR transcripts reliably because it relies on read alignments which fail for the highly repetitive and rearranged IG heavy (IGH), IG light (IG kappa (IGK) and lambda (IGL)), TCR alpha (TRA), and beta (TRB) loci. To capture AIR transcripts reliably, we first identified R2C2 reads which aligned to the constant region exons in the IG and TR loci. We then determined

which of these reads contained a high quality V segment using IgBlast [37]. Finally, we used these filtered reads to determine consensus sequences for each locus and cell (Fig. 15A).

For many B cells we determined multiple sequences for different isotypes (IGHM, IGHD, IGHG (1, 2, 3,and 4), and IGHA (1 and 2) and isoforms (membrane bound and secreted). In the vast majority of cases (103/108) (Fig. 15B), transcripts contained the same V segment, indicating that they represent alternative splicing products of the same rearrangement. We succeeded in determining paired IG sequences for 110 B cells and 381 T cells which represent 61% and 17% of all B and T cells analyzed in this study, respectively (Fig. 15C). Importantly, as would be expected for a random sample of B cells, the V(, D,) and J segment usage composition of the paired transcripts of these cells was highly diverse (Fig. 15C)

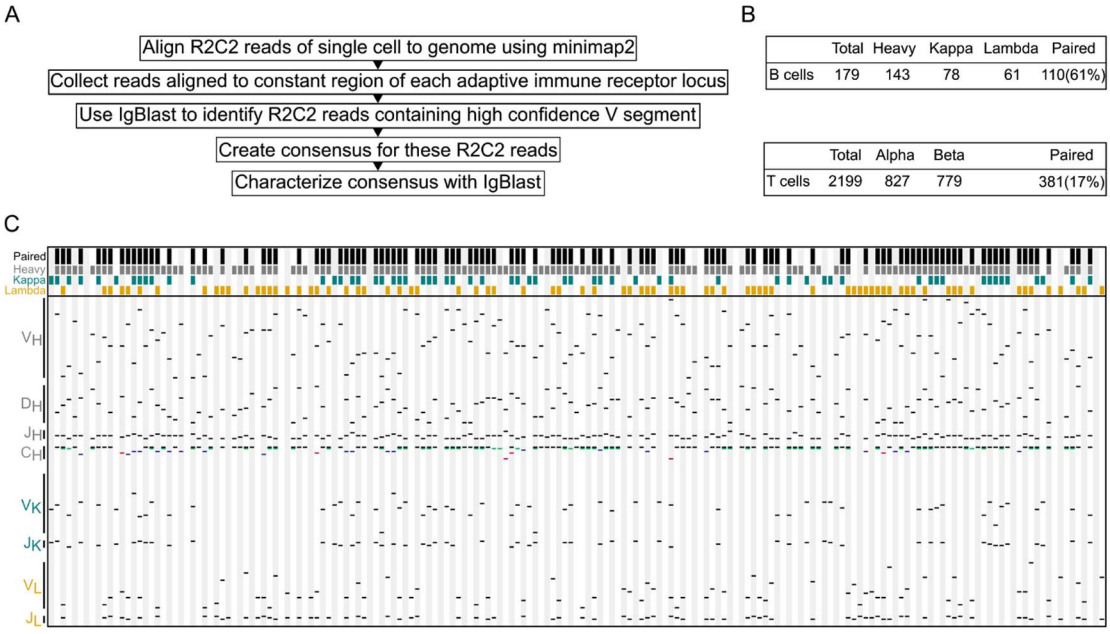

Fig. 15: IG and TCR transcripts can be identified and paired in 10X R2C2 data. *A) The workflow to identify antibody (IG) and T cell receptor (TCR) transcripts for each individual cell. B) Numbers of cells for which IG or TCR transcripts could be identified and paired. C) Schematic of IG identification, composition, and pairing. Each column represents a single B cell. Colored blocks on top of each column indicate whether a cell contains paired IG transcripts (black), whether an IGH (Heavy: grey), IGK (Kappa: teal), or IGL (Lambda: orange) transcripts was detected. Below the diversity of the detected sequences is shown. Black lines indicate which gene segments were used when an IG sequence was recombined from the germline genome. In C_H, it is also shown which isotype(s) we detected (IGHM: black, IGHD: green, IGHA1 or 2: red, IGHG1-4: blue) for each cell.*

Discussion

Here, we present a method to analyze highly-multiplexed full-length single-cell transcriptomes that does not require short-read sequencing. We processed 10ng of cDNA generated as an intermediate product of the 10X Genomics Chromium Single Cell 3' Gene Expression Solution into R2C2 sequencing libraries. We sequenced these libraries and demultiplexed the resulting data to produce over 12 million unique transcript molecules generated from ~3000 PBMCs. We used these single cell data to determine monocyte, T cell, and B cell clusters, generate isoform-level transcriptomes for these cell types, investigate single-cell isoform diversity, and pair adaptive immune receptor transcripts.

The ability to analyze the full-length transcriptomes of single cells without the need for Illumina short-read data has the potential to simplify experimental workflows. The ability to perform this analysis on low cost ONT sequencers will make it more accessible. This is made possible through the use of the R2C2 sample preparation method which can increase the base accuracy of ONT MinION sequencers to ~99%. In this study, the R2C2 base accuracy was closer to 98% due to shorter raw reads. We aimed for shorter raw reads to increase R2C2 read numbers and, to this end, reduced the stringency of our size-selection prior to sequencing.

Outside of R2C2, raw nanopore reads are becoming more accurate and are used to analyze 10X cDNA with the help of Illumina data or by themselves using modified 10X protocols with longer indexes. Further, single cell studies using the PacBio Sequel II, while limited in overall throughput and hampered by per-read cost of the sequencer, benefit from the very high accuracy of the reads which simplifies computational analysis. Going forward, the trade-off between throughput, cost, and accuracy of ONT MinION and PromethION as well as PacBio Sequel II sequencers will have to be considered closely and the best compromise may well vary between studies.

At current throughput and accuracy, the combination of ONT sequencers and the R2C2 method allows the analysis of thousands of cells. An increase in read output will make it possible to either analyze more cells or sequence all transcripts reverse transcribed by the 10X Genomics workflow. In this current study, with about 3,000 R2C2 reads per cell, we captured about 67% of the molecules present in an exhaustively sequenced Illumina dataset of the same cDNA. This was sufficient to cluster cell types and generate single-cell transcriptomes. An increase in accuracy would make future demultiplexing and UMI merging steps more efficient. While our demultiplexing strategy can handle sequencing errors (see Methods), at 98% accuracy it still only manages to demultiplex ~81% of R2C2 reads, which is better than previously published approaches, but not ideal[67,71]. Increasing accuracy could increase this number significantly. Paired with higher throughput, future experiments could only retain UMIs which were observed more than once, similar to how we analyze Illumina data (see Methods).

Beyond establishing this method, we generated high-quality transcriptomes for Monocyte, B cell, and T cell populations. Because the majority of PBMCs are T cells, the T cell transcriptome is the most comprehensive of those three and should serve as a resource for understanding the biology of these adaptive immune cells.

We then used a framework developed for a previous study[70] to show that these cell types show differential isoform expression. The ability to identify differentially expressed isoforms expands the quality of information that can be extracted from single-cell experiments and opens the door to a much more nuanced understanding of gene regulation.

Beyond investigating isoform expression on the cell type level, we investigated the extent of isoform diversity on the single-cell level. While some genes showed low isoform diversity, i.e. most cells express the same isoform, some genes showed high diversity, i.e. many cells express unique isoforms. This wide range of isoform diversity will pose a formidable challenge for single-cell level differential isoform expression analysis going forward. Future studies into how this wide range of isoform diversity is maintained and used by cells are bound to generate fascinating insights into transcript processing and cellular function.

In the meantime, using isoforms identified independently for single cells can already inform isoform identification. While different isoform identification tools like TALON[81], FLAIR[82], or StringTie2[82], and Mandalorion use different strategies when identifying and filtering isoforms, they all rely on some form of read coverage cut-off to differentiate real isoforms from the noise produced by any sequencing method. However, PCR or sequencing artifacts generated within a single cell can overcome these cut-offs and result in the false-positive identification of isoforms. The information of how many single cells express an isoform could therefore aid in the identification of real or biologically meaningful isoforms as each single cell can be seen as an independent biological replicate.

Finally, taking advantage of the single-cell nature of this dataset, we performed analysis on the most complex part of T cell and B cell transcriptomes, namely adaptive immune receptor transcripts. By sequencing and pairing adaptive immune receptor transcripts expressed by single T and B cells, we showcased the power of long reads for resolving even the most challenging transcript isoforms – without the need for specialized

protocols. This will be of particular use when analyzing complex samples that contain, but aren't limited to, immune cells like solid or liquid tumors.

Methods

Single cell cDNA library preparation

Full-length cDNA pools and Illumina libraries were prepared by 10X Genomics. PBMCs were sourced from Stemcell Technologies and prepared for sequencing using the 10X Genomics Chromium Single Cell 3' Gene Expression Solution. Preparation of the cDNA was done according to manufacturer's instructions with the exception of the extension time for the final PCR reaction which was standard 1 minute for replicate 1 but increased to 4 minutes for replicate 2.

Illumina sequencing and read processing

Illumina libraries were sequenced on the Illumina NextSeq with Read1 = 26bp and Read2 = 134bp.

Overall a NextSeq flowcell generated 107,911,006 reads for replicate 1 and 75,753,410 reads for replicate 2. Reads were then demultiplexed and collapsed by determining the 1500 most frequent cellular barcodes, perfectly matching cell barcodes to the most frequent, and then filtering for unique cell barcode/10X UMI combinations.

Reads for each cell were then aligned to the human genome (hg38) using `STAR` (*--runThreadN 30 --genomeDir /path/to/STAR/index/ --outSAMtype BAM SortedByCoordinate --readFilesIn /path/to/reads --outFileNamePrefix /path/to/alignment/dir*).

Nanopore sequencing and read processing

Full-length cDNA pools were prepared as described previously. In short, 10ng of cDNA is circularized using a DNA splint compatible with 10X cDNA and the NEBuilder HIFI DNA Assembly Master Mix (NEB). The DNA splint was generated by primer extension of the following oligos:

```
>10X_UMI_Splint_Forward (Matches 10X PCR primer)
AGATCGGAAGAGCGTCGTGTAG
TGAGGCTGATGAGTTCCATANNNNNTATATNNNNNATCACTACTTAGTTTTTTGATAGCTTCAAGCCA
GAGTTGTCTTTTTCTCTTTGCTGGCAGTAAAAG
>10X_UMI_Splint_Reverse (Matches ISPCR Primer)
CTCTGCGTTGATACCACTGCTT
AAAGGGATATTTTCGATCGCNNNNNATATANNNNNTTAGTGCATTTGATCCTTTTACTCCTCCTAAAG
AACAACCTGACCCAGCAAAAGGTACACAATACTTTTACTGCCAGCAAAGAG
```

Non-circularized DNA is digested using Exonucleases I, III, and Lambda. Circularized DNA is amplified using rolling circle amplification using Phi29 (NEB). The resulting HMW DNA is debranched using T7 Endonuclease (NEB) and purified and size-selected using SPRI beads. This DNA containing concatemers of the originally circularized cDNA is then sequenced using the LSK-109 kit on either ONT MinION or PromethION sequencers. The resulting raw reads were processed into consensus reads using the C3POa pipeline. All consensus reads were then assigned a cell of origin. In a first step, we determined the most common ~1500 cellular identifiers in our sample using a simple counting strategy. Then, we assigned reads to the most similar cellular identifiers if they fit the following criteria:

1.) L1 < 3

and

2.) L1 < L2 - 1

where

L1 is the Levenshtein distance between the read's cellular identifier and the most similar known cellular identifier

and

L2 is the Levenshtein distance between the read's cellular identifier and the second most similar known cellular identifier.

These consensus reads were demultiplexed based on their cell assignment, they were merged if the contained the similar UMIs in their splint back-bones using the `ExtractUMIs` and `MergeUMIs` utilities (https://github.com/rvolden/10xR2C2). The resulting reads were then merged again if the contained the similar 10X UMIs in their adapters using the `ExtractUMIs` and `MergeUMIs` utilities (https://github.com/rvolden/10xR2C2).

The resulting Splint/10X-UMI merged R2C2 consensus reads were then demultiplexed based on their initial cell assignments. If a Splint/10X-UMI merged R2C2 consensus read was generated by merging reads with different cell assignments it was discarded. Reads for each cell were then aligned to the human genome (hg38) using `minimap2[29]` (*-ax splice --secondary=no -G 400k*).

Cell type clustering

Both Illumina and R2C2 data were analyzed in the same way independently. First gene expression tables were generated using `featureCounts`[75]. Then these tables were parsed for input into the Seurat R package (v3)[83]. Seurat generated cell type clusters using the

following main settings (*min.cells=3, min.features=200, percent.mt<5, 2500>nFeature_RNA>200, nfeatures=2000, dims=1:10, resolution=0.08 (0.08 used for nanopore, 0.03 for Illumina), log normalization, and vst selection*).

For each cell, cell type information was extracted based on location for downstream analysis.

Isoform analysis

We generated high confidence isoforms using the latest version of the Mandalorion pipeline (Episode III.5, https://github.com/rvolden/Mandalorion).

Cell type transcriptomes:

All reads and subreads assigned to cells of a cell type were pooled. Mandalorion was run on these files with the following settings:

```
-c /path/to/config_file
-m /path/to/NUC.4.4.mat
-I 300
-g /path/to/gencode.v37.annotation.gtf
-G /path/to/hg38.fa
-a /path/to/10x_Adapters.fasta
-f /path/to/Pooled_reads.fa
-b /path/to/Pooled_subreads.fa
-p /path/to/output_folder
-e ATGGG,AAAAA
```

with `10x_Adapters.fasta` containing the following sequences:

```
>3Prime_adapter
CTACACGACGCTCTTCCGATCT
>5Prime_adapter
AAGCAGTGGTATCAACGCAGA
```

Single-cell transcriptomes:

Mandalorion was run on the reads, read alignments, and subreads of each individual cell.

Mandalorion was run with the following settings:

```
-c /path/to/config_file
-I 300
-g /path/to/gencode.v37.annotation.gtf
-G /path/to/hg38.fa
-a /path/to/10x_Adapters.fasta
-f /path/to/SingleCell_reads.fa
-b path/to/SingleCell_subreads.fa
-p path/to/output_folder
-e ATGGG,AAAAA
-R 2
```

Note that we reduced the minimum number of reads required to identify an isoform to 2.

The resulting isoform psl files were converted to gtf files and classified using the `sqanti_qc.py` program and the following settings:

```
-g
-n
-t 24
-o output_prefix
-d /path/to/output_folder
/path/to/gtf_file
/path/to/gencode.v37.annotation.gtf
/path/to/hg38.fa
```

Isoform diversity analysis

Similar isoforms were merged using the `merge_psls.py` utility which accepts a list of isoform fasta and psl files and merges isoforms if they:

1. Use all the same splice sites

   This step is base-accurate but will treat splice site a single base pair apart as equivalent if one site is much less abundant than the other

2. Use the similar start and end sites

   This step will consider sites similar if they are at most 10nt apart. Because isoforms are iteratively grouped at this step, individual isoforms in a merged group might have sites that are further than 10nt apart but are connected by a third isoform between them.

Adaptive Immune receptor analysis

For each cell, reads aligning to the T cell or B cell receptor loci were extracted from sam files using `samtools view`[84] and the below genomic coordinates.

```
IGH: chr14: 105,533,853 - 106,965,578
IGK: chr2:   89,132,108 -  90,540,014
IGL: chr22:  22,380,156 -  23,265,691
TRA: chr14:  22,178,907 -  23,021,667
TRB: chr7:  141,997,301 - 142,511,567
```

Reads were then analyzed for each cell and locus (and for IGH, each isotype/isoform) separately by filtering reads for a high-quality match to a V segment retrieved from IMGT[84] using `IgBlast`[85] and the following settings:

```
-germline_db_V /path/to/V_segments
-germline_db_J /path/to/J_segments
-germline_db_D /path/to/D_segments
-organism human
-query /path/to/reads.fasta
[-ig_seqtype TCR ] - only for T cell receptors
-auxiliary_data optional_file/human_gl.aux
-show_translation
-outfmt 19
```

Filtered reads for each cell were then used to generate consensus reads for each locus. Those consensus reads were then assigned V, (D,) and J segments using `IgBlast` and the same settings as above. All scripts used for this analysis and a wrapper script automating this analysis are available at https://github.com/christopher-vollmers/AIRR-single-cell.

Data Access

We uploaded all data generated for this study to the SRA where it is available under BioProject accession PRJNA599962.

B cell, T cell, and Monocyte transcriptomes are available at https://users.soe.ucsc.edu/~vollmers/10XR2C2/.

Code Access

We have made the code required to demultiplex R2C2 reads and format gene expression matrices for Seurat available on GitHub (https://github.com/rvolden/10xR2C2). Code for AIRR analysis is also available on GitHub (https://github.com/christopher-vollmers/AIRR-single-cell).

Conclusion

The work presented here is a showcase of how to overcome challenges in obtaining highly accurate full-length cDNA sequences at a reasonable cost. The R2C2 and C3POa methods allow for high throughput full-length cDNA isoform sequencing in single cells. The R2C2 method is a reliable method for capturing full-length cDNA molecules without significant bias. By increasing nanopore sequencing base accuracy from 85% to 95%, it was possible to sequence a handful of single cells. With further improvements to base ONT accuracy, R2C2 library preparation, and consensus calling with C3POa, we were able to increase accuracy from 95% to over 99%. Improving the base accuracy is vital for increasing the method's capacity for highly multiplexed experiments, such as single-cell cDNA sequencing with droplets.

Using R2C2 and C3POa, we demonstrated that it was possible to sequence highly multiplexed single-cell cDNA samples without needing to lean on short-reads to reliably demultiplex long-reads into cells. Using our single-cell cDNA data, we were able to determine B cell, T cell, and monocyte populations based on their gene expression. Using the cluster information, we were able to make isoform-level transcriptomes for each cell type as well as investigate differential isoform usage between cell types. Additionally, we were able to use the single-cell data to investigate single-cell isoform diversity and pair together adaptive immune receptor transcripts. Using our methods for full-length single-cell sequencing has the potential to simplify existing single-cell isoform sequencing protocols because it only uses one sequencing technology while retaining cost effectiveness. Our methods are also well-suited for sequencing adaptive immune receptor transcripts without needing a targeted sequencing protocol.

While there are still optimizations to be made at many points in these methods, I believe that we are currently at the point where we can effectively use nanopore sequencing to study isoform level transcriptomes in single cells. Our lab has also demonstrated that

R2C2 and C3POa are highly powerful tools for everything from polar bear transcriptome analysis[69] to *Drosophila Melanogaster* genome assembly[86]. In the future, I would expect to see these methods used by labs that are interested in cheaply doing isoform-level analyses.

References

1. Adams, M. D. *et al.* Sequence identification of 2,375 human brain genes. *Nature* **355**, 632–634 (1992).
2. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
3. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680 (1996).
4. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
5. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756-63 (2014).
6. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
7. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
8. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
9. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
10. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
11. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
12. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16**, 150 (2015).
13. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
14. Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* **16**, 204 (2015).

15. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
16. Treutlein, B., Gokce, O., Quake, S. R. & Südhof, T. C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1291-9 (2014).
17. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
18. Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4821-30 (2013).
19. Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708 (2016).
20. Tilgner, H. *et al.* Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3* **3**, 387–397 (2013).
21. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9869–9874 (2014).
22. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
23. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
24. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4259.
25. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9726–9731 (2018).
26. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
27. Stepanauskas, R. Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15**, 613–620 (2012).
28. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).

29. Zhang, N. & Bevan, M. J. CD8(+) T cells: foot soldiers of the immune system. *Immunity* **35**, 161–168 (2011).

30. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

31. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742 (2015).

32. Tilgner, H. *et al.* Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* **28**, 231–242 (2018).

33. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).

34. Kuo, R. I. *et al.* Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323 (2017).

35. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6**, 31602 (2016).

36. Li, C. *et al.* INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5**, 34 (2016).

37. Sotillo, E. *et al.* Convergence of Acquired Mutations and Alternative Splicing of CD19 Enables Resistance to CART-19 Immunotherapy. *Cancer Discov.* **5**, 1282–1295 (2015).

38. Fischer, J. *et al.* CD19 Isoforms Enabling Resistance to CART-19 Immunotherapy Are Expressed in B-ALL Patients at Initial Diagnosis. *J. Immunother.* **40**, 187–195 (2017).

39. Sommermeyer, D. *et al.* Fully human CD19-specific chimeric antigen receptors for T-cell therapy. *Leukemia* **31**, 2191–2199 (2017).

40. Cole, C., Byrne, A., Beaudin, A. E., Forsberg, E. C. & Vollmers, C. Tn5Prime, a Tn5 based 5' capture method for single cell RNA-seq. *Nucleic Acids Res.* **46**, e62 (2018).

41. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

42. Lee, C., Grasso, C. & Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464 (2002).

43. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

44. Tardaguila, M. *et al.* SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *bioRxiv* 118083 (2017) doi:10.1101/118083.

45. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

46. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

47. Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43**, W580-4 (2015).

48. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* **41**, W597-600 (2013).

49. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

50. Li, H. Minimap2: fast pairwise alignment for long nucleotide sequences. arXiv. (2017).

51. Lindenbaum, P. JVarkit: java-based utilities for Bioinformatics. (2015) doi:10.6084/m9.figshare.1425030.v1.

52. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

53. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).

54. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).

55. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).

56. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

57. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).

58. Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15**, 563–565 (2018).

59. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).

60. Miragaia, R. J. *et al.* Single-Cell Transcriptomics of Regulatory T Cells Reveals Trajectories of Tissue Adaptation. *Immunity* **50**, 493-504.e7 (2019).

61. Van Hove, H. *et al.* A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment. *Nat. Neurosci.* **22**, 1021–1035 (2019).

62. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

63. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

64. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).

65. Cole, C., Byrne, A., Adams, M., Volden, R. & Vollmers, C. Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *bioRxiv* 761437 (2019) doi:10.1101/761437.

66. Rebboah, E. *et al.* Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. 56 (2021).

67. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat. Commun.* **11**, 4025 (2020).

68. Philpott, M. *et al.* Highly accurate barcode and UMI error correction using dual nucleotide dimer blocks allows direct single-cell nanopore transcriptome sequencing. *bioRxiv* 2021.01.18.427145 (2021) doi:10.1101/2021.01.18.427145.

69. Byrne, A. *et al.* Depletion of Hemoglobin Transcripts and Long-Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (Ursus maritimus). *Front. Genet.* **10**, 643 (2019).

70. Vollmers, A. C., Mekonen, H. E., Campos, S., Carpenter, S. & Vollmers, C. Generation of an Isoform-level transcriptome Atlas of Macrophage Activation. *Journal of Biological Chemistry* 100784 (2021) doi:10.1016/j.jbc.2021.100784.

71. Singh, M. *et al.* High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.* **10**, 3120 (2019).

72. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

73. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

74. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

75. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

76. Leduc, I., Preud'homme, J. L. & Cogné, M. Structure and expression of the mb-1 transcript in human lymphoid cells. *Clin. Exp. Immunol.* **90**, 141–146 (1992).

77. Schanberg, L. E., Fleenor, D. E., Kurtzberg, J., Haynes, B. F. & Kaufman, R. E. Isolation and characterization of the genomic human CD7 gene: structural similarity with the murine Thy-1 gene. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 603–607 (1991).

78. Auron, P. E. *et al.* Nucleotide sequence of human monocyte interleukin 1 precursor cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 7907–7911 (1984).

79. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* (2018) doi:10.1101/gr.222976.117.

80. Wyman, D. *et al.* A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv* 672931 (2020) doi:10.1101/672931.

81. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* 410183 (2018) doi:10.1101/410183.

82. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

83. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

84. Lefranc, M.-P. *et al.* IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol.* **4**, 17–29 (2004).

85. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34-40 (2013).

86. Adams, M. *et al.* One fly-one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster. *Nucleic Acids Res.* **48**, e75 (2020).