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[1,2]. Pairing transcriptome analysis with the high-throughput processing of single cells has provided unprecedented insight into cellular heterogeneity[3,4]. Among many other studies, researchers have leveraged the strengths of high-throughput single-cell transcriptome analysis to create single cell maps of the mouse[5,6] or C. elegans[7] model organisms, to elucidate a new cell type in the lung involved in cystic fibrosis[8], and to increase our knowledge of adaptive and innate immune cells[9–12].

High-throughput single-cell transcriptome analysis however comes with trade-offs. In particular, droplet- or microwell-based methods like Drop-seq[13], 10X Genomics[14], and Microwell-Seq[6] or Seq-Well[15] 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[16–19]. These methods have now been used to analyze single cell cDNA pools generated by different methods, both well-[20,21] and droplet-based[22,23], 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[23]. 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[19,20,25]. 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 generate 1500 cells each for two replicates. We divided the full-length cDNA intermediate generated by the standard 10X Genomics protocol to perform both short- and long-read sequencing (Figure 1A).

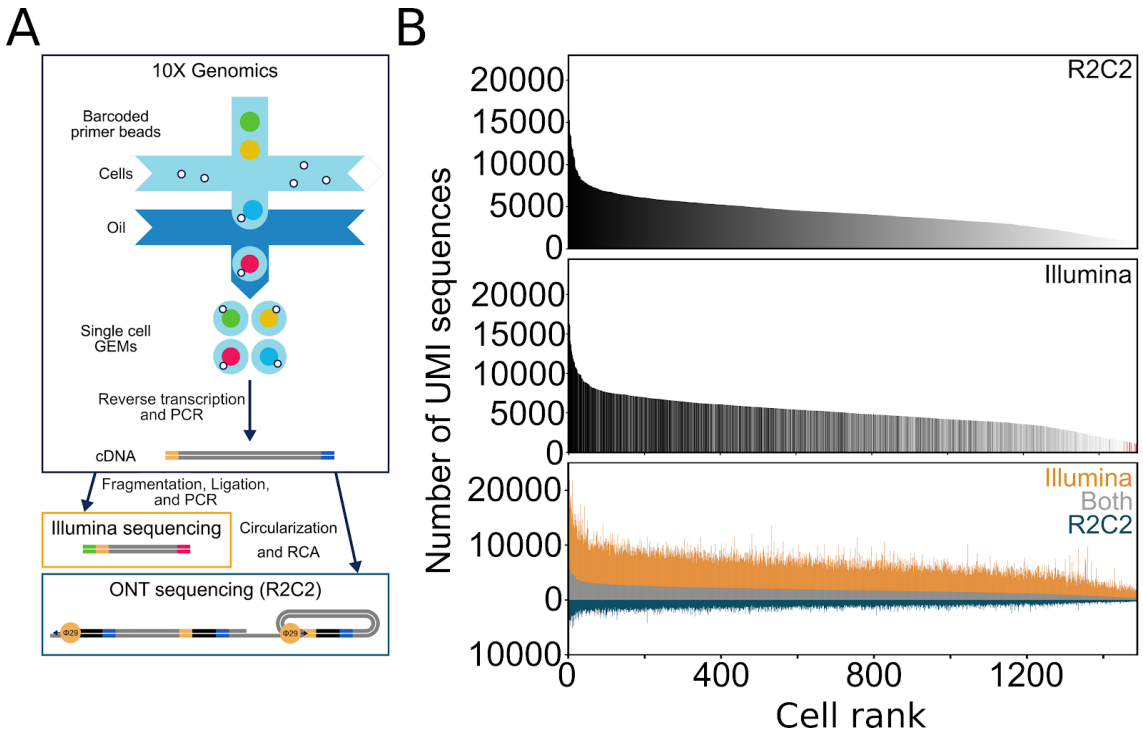


Fig. 1: Data Generation and Characteristics. A) Thousands of peripheral blood mononuclear cells (PBMCs) were processed using the 10X Genomics Chromium Single Cell 3' Gene Expression Solution. The resulting full-length cDNA was either fragmented for Illumina sequencing or processed using the R2C2 workflow.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

For sequencing on the Illumina NextSeq, we fragmented the full-length cDNA according to the standard 10X protocol. We demultiplexed and combined the resulting reads based on cellular barcodes and unique molecular identifiers (UMIs) associated with every amplified transcript molecule during reverse transcription (see Methods). In this way, 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).

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 S1). We further combined 7-10% of consensus reads because unique molecular identifiers (UMIs) embedded in the dsDNA splint used for circularization identified them as originating from the same cDNA molecule. This resulted in highly accurate merged consensus reads (Table S2). Overall, this process generated ~16.8 million R2C2 reads (Table 1).

	Basecalled reads	Post-processed R2C2 reads	Splint-UMI merged R2C2 reads	Demultiplexed R2C2 reads	10X-UMI merged R2C2 reads (Unique transcripts)
Replicate 1	29,527,932	9,480,134 (32.1%)	8,833,532 (93.17%)	6,394,275 (72.39%)	4,754,496 (74%)
Replicate 2	26,526,607	8,861,132 (33.4%)	7,942,929 (89.65%)	5,737,067 (72.23%)	4,198,397 (73%)

Table 1: Read numbers throughout processing.

Next, we demultiplexed these reads based on the 10X cellular barcodes they contained. 72% of R2C2 reads covering 10X molecules 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 a previous studies[23,26]. Moreover, 99.0% of the 3000 cellular identifiers we determined independently from the R2C2 data set also appeared in the Illumina data set. Further, the distribution of reads between cells was also highly similar between the data sets (Fig 1B). After demultiplexing, ~12 million R2C2 reads were merged if they contained perfectly matching 10X UMIs resulting in ~9 million full-length sequences originating from unique transcripts across both replicates with a median sequence accuracy of 96.4%. As shown in previous studies analyzing 10X cDNA with long reads[23,24], R2C2 reads appeared to cover entire transcripts (Fig. 2)

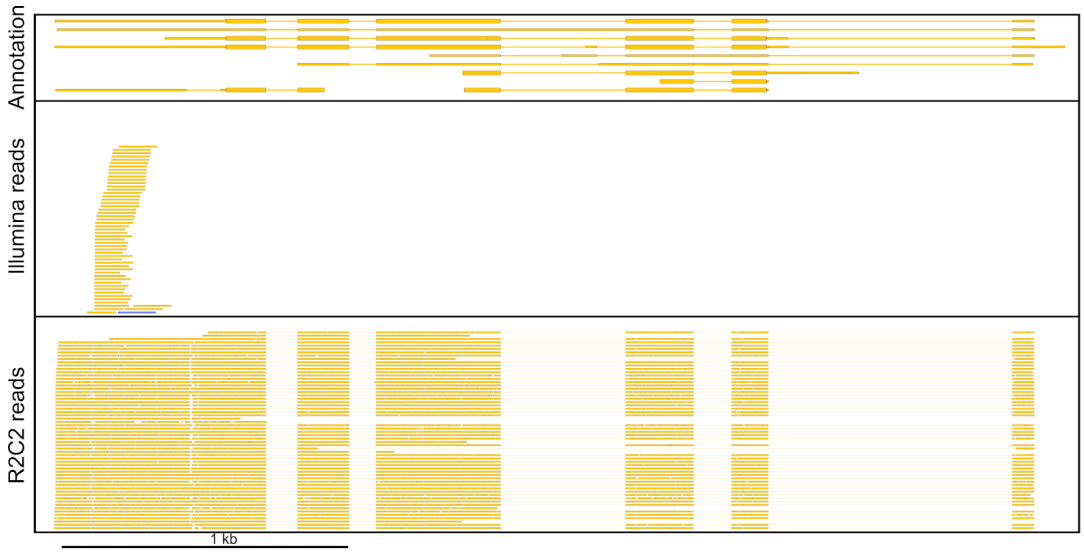


Fig. 2: R2C2 reads sequence 10X full-length cDNA transcripts. Genome Browser shots of ACTB. Genome annotation is shown on top and Illumina reads (center) R2C2 reads (bottom) aligning to the locus are shown below. Both Illumina and R2C2 read alignments were randomly subsampled to 60 reads. The directionality of features is indicated by color ("top strand"=blue, "bottom strand"=yellow). Data for replicate 1 are shown.

For each cell, 55% of the UMIs captured in the R2C2 data set were also captured by the Illumina data set, which at a coverage of >10 reads per molecule (202,469,707 reads / 15,264,862 molecules) can be assumed to be nearly comprehensive. Because we would expect the comprehensive Illumina data set to cover all UMIs in the R2C2 data set, the remaining ~4% error in R2C2 reads likely causes the remaining 45% of UMIs to be unmatched. Indeed, at a depth of 6 million reads per replicate, a 4% error rate, and a 10nt UMI, we roughly expect $6x10^{6}*(1-0.96^{10})$ or 2 million (33%) reads to contain at least one sequencing error in their UMI. As such, base-perfect UMI identification in the R2C2 data set seems to be about as accurate as the residual error in R2C2 reads allows. Importantly, although their transcript molecule of origin cannot be unambiguously identified, R2C2 reads with UMIs containing 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[27,28]. For comparison, Illumina reads generated from the same cDNA were aligned using STAR and also processed using featureCounts[29]. Median Pearson-r values for R2C2 and Illumina-based gene expression for the same cell showed high correlation at 0.74 (Fig. S1).

We then clustered R2C2 and Illumina data sets independently using the Seurat analysis package[30]. R2C2 and Illumina data sets both grouped into three cell-type clusters. Based on marker gene expression, the major cell-types could be identified as B cells (CD79A)[31], T cells (CD7)[32], and Monocytes (IL1B)[33] – the expected composition of a PBMC sample (Fig. 3, S2). Importantly 99.5% of cells that were clustered in both data sets associated with the same cell-type in the two data sets.

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.

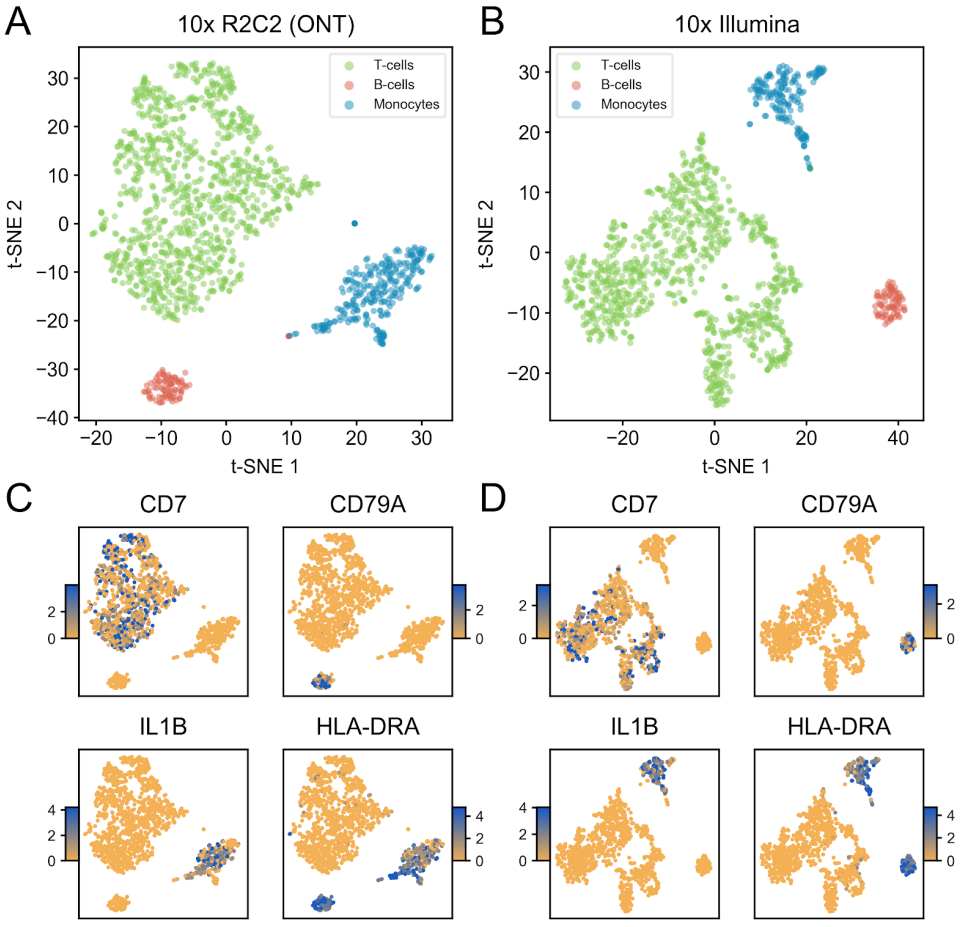


Fig. 3: R2C2 and Illumina data sets independently cluster into B cells, T cells, and Monocytes.
Gene expression profiles were determined independently for each cell in R2C2 and Illumina data sets. The Seurat package was then used to cluster cells based on the gene expression profiles. The cells in R2C2 (A) and Illumina (B) data sets 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. First, as previously established[24], we pooled all reads associated with the cells of each cell type and then identified transcript isoforms for each cell-type using Mandalorion[19–21,25]. The majority (65-70%) of isoforms generated by Mandalorion for the individual cell-types were classified by SQANTI[34] as either 'full-splice-match' or 'novel-in-catalog' which represent likely full-length isoforms. 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 2).

Cell-type	Number of cells	Number of reads	Number of genes with Isoforms	Number of Isoforms
B cells	187	509,274	2,782 (plus 1,019 novel genes)	5,554
T cells	2,211	6,805,517	7,696 (plus 1,854 novel genes)	21,810
Monocytes	554	1,572,175	4,180 (plus 1,320 novel genes)	9,622

Table 2: Cell-type specific full-length transcriptome characteristics

Isoform diversity is highly variably 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 data set we present here, we have sufficient coverage to generate isoforms for each cell independently. Using Mandalorion, we generated a median of 160 isoforms per cell. We then analyzed isoforms across all cells in a cell-type. To this end, we merged identical isoforms expressed by different cells. We then determined how many cells expressed any given gene and isoform.

Interestingly, isoform diversity varied greatly between genes (Fig. 4A). On one end of the spectrum, genes encoding ribosomal proteins in particular are expressed in the majority of cells in each cell-type, yet we identify few unique isoforms for these genes. For example, 1255 cells expressed a total of 1258 isoforms (as determined by Mandalorion) of the ribosomal protein gene RPL35. After merging all identical isoforms, only 7 unique isoforms remained. 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, 906 cells expressed a total of 1943 LMNA isoforms. After merging all identical LMNA isoforms, 691 unique isoforms remained.

To quantify this range in isoform diversity systematically, we calculated the ratio of unique isoforms we identify for a gene to the number of cells expressing it. By calculating this unique-isoforms-per-cell (UIPC) ratio for all genes expressed by at least 10% of cells in a cell-type we found a wide range of isoform diversity (Fig. 4B). In all cell-types, genes encoding ribosomal proteins represent the genes with the lowest isoform diversity (Table S3) with many of these genes showing a UIPC ratios close to 0. The LMNA gene has the highest or second highest isoform diversity in all cell-types with a UIPC ratios between 0.8 and 1.2.

To visualize the isoform composition of genes with a range of UIPC ratios, we extracted the isoforms expressed by 25 random cells for RPL35, CD37, NFKB1D, and LMNA (Fig. 4C). As expected, all 25 cells expressing the RPL35 genes expressed a single virtually identical isoform. Cells expressing CD37 appear to express two majority isoforms but also diverse other isoforms. Most cells expressing NFKBID appeared to express a different unique isoform while the cells expressing LMNA often expressed several unique isoforms.

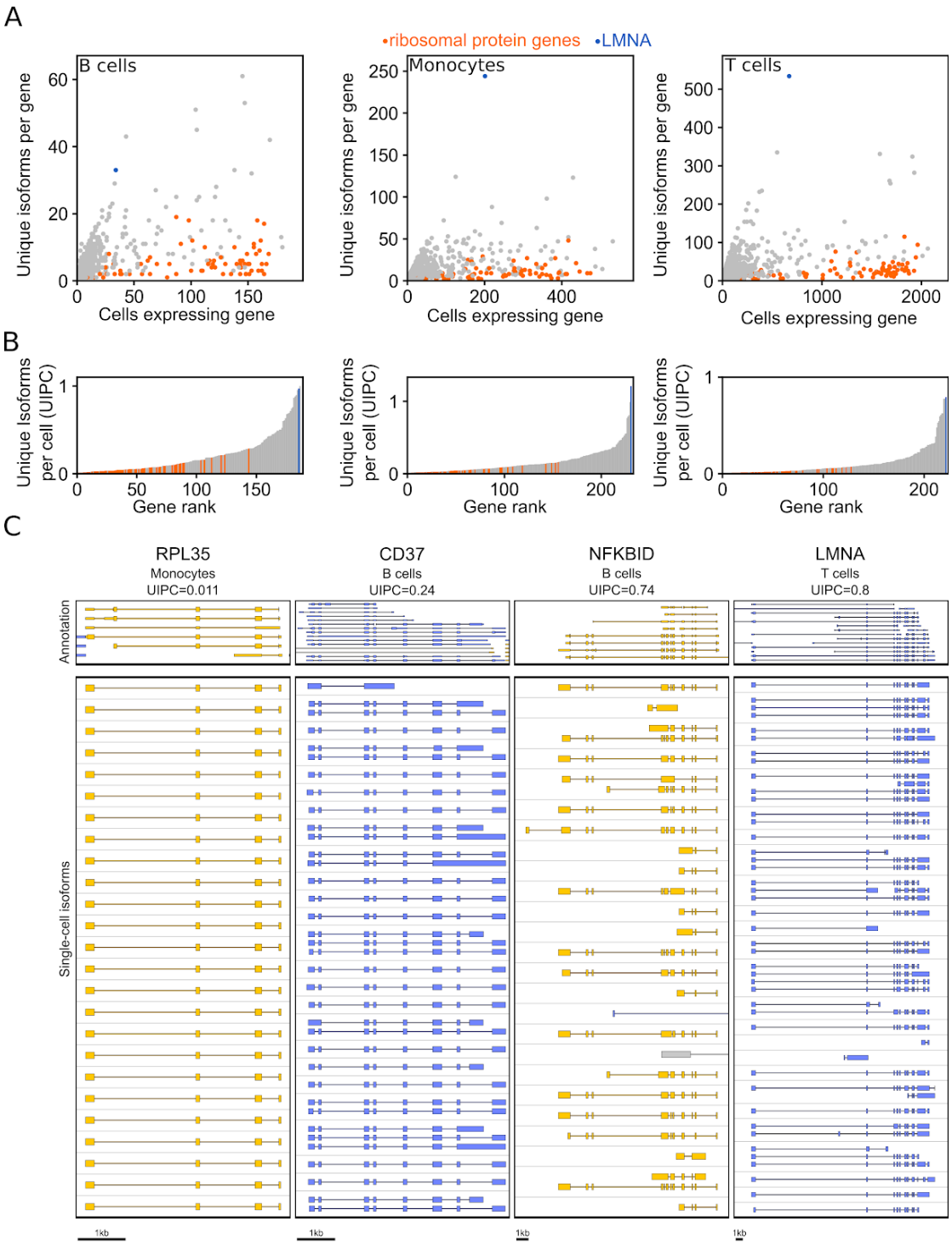


Fig. 4: Genes show a wide range of isoform diversity.
We generated an isoform level transcriptome for each cell in our data sets and then analyzed the isoform diversity for different genes. A) The correlation of the number of cells expressing an isoform for a gene and how many unique isoforms we identified for that gene is shown as a scatter plot for the indicated cell-types. Genes encoding ribosomal proteins and LMNA proteins are shown in orange and blue respectively. B) The number of unique isoforms of a gene per cells expressing that gene was calculated for genes that were expressed by at least 10% of cells in the indicated cell-type. The resulting Unique Isoforms per cell (UIPC) measure is shown as a rank ordered histogram. C) Genome Browser shots of genes representing different UIPCs are shown. Genome annotation is shown on top and Isoforms expressed by 25 random single cells shown below separated by grey lines. The Directionality of features is indicated by color (“top strand”=blue, “bottom strand”=yellow)

Extracting paired adaptive immune receptor sequences from B and T cells

We investigated whether our data sets 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[35]. Finally, we used these filtered reads to determine consensus sequences for each locus and cell (Fig. 5A).

For many B cells we determined multiple sequences for different isotypes (IGHM, IGHD, IGHG(1, 2, 3,and 4), and IGHA(1 and 2) (Table S4) and isoforms (membrane bound and secreted). In the vast majority of cases (92/97) (Fig. 5B), 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 ~100 B cells and ~200 T cells which represent 52% and 10% of all B and T cells analyzed in this study, respectively (Fig. 5C). 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. 5C)

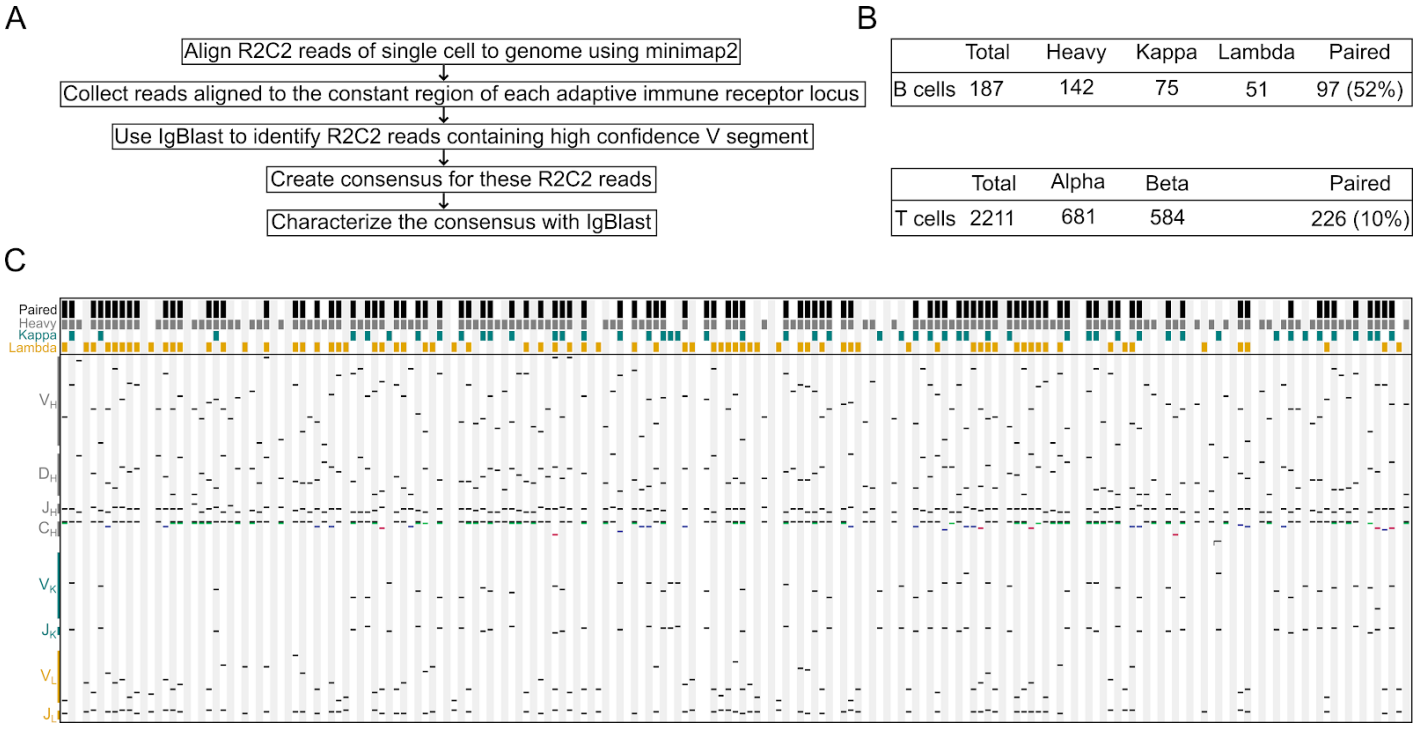


Fig. 5: IG and TCR transcripts can be identified and paired in 10X R2C2 data. A) The workflow to identify antibody (IG) and T cell receptor (TCR) transcripts for each individual cell. B) Numbers of cells for which IG or TCR transcripts could be identified and paired. C) Schematic of IG identification, 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 9 million unique transcripts 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 clusters, 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 ~98%. In this study, the R2C2 base accuracy was closer to 96% 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 and used the ONT PromethION sequencer which inherently seems to produce slightly shorter raw sequencing reads (Table S1). Going forward, the trade-off between throughput, cost, and accuracy of ONT MinION and PromethION 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 estimated that we captured about 60% (based on ~5000 molecules per cell in Illumina data set) of all reverse transcribed molecules. 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 96% accuracy it still only manages to demultiplex ~72% of R2C2 reads, which is better than previously published approaches, but not ideal[23,26]. Further, at 96% accuracy, about 33% of reads will contain at least one error in their 10nt 10X UMI. Increasing accuracy could reduce 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 performed additional 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 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 analysing complex samples that contain, but aren't limited to, immune cells like solid or liquid tumors.

Finally, we performed initial analysis into isoform diversity which varied widely between genes. While some genes showed low isoform diversity, i.e. most cells express the same isoform, some genes showed high diversity, i.e. most cells express one or more unique isoforms. This wide range of isoform diversity will pose a formidable challenge for isoform level differential 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 cellular function.

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
TGAGGCTGATGAGTTCCATANNNNNTATATNNNNNATCACTACTTAGTTTTTTGATAGCTTCAAGCCAGAGTTGTCTTTTTCTCTTTGCTGGCAGTAA
AAG
>10X_UMI_Splint_Reverse (Matches ISPCR Primer)
CTCTGCGTTGATACCACTGCTT
AAAGGGATATTTTCGATCGCNNNNNATATANNNNNTTAGTGCATTTGATCCTTTTACTCCTCCTAAAGAACAACCTGACCCAGCAAAAGGTACACAAT
ACTTTTACTGCCAGCAAAGAG
```

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 (Table S1). The resulting raw reads were processed into consensus reads using the C3POa pipeline. These consensus reads are then merged if the contained the similar UMIs in their splint back-bones using the `ExtractUMIs` and `MergeUMIs` utilities (https://github.com/rvolden/10xR2C2). All consensus reads were then demultiplexed. 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.

Once demultiplexed, reads for each cell were merged again using the `MergeUMIs10x` utility (https://github.com/rvolden/10xR2C2), this time based on the UMI present in the 10X oligodT primer. Here, we only considered perfect UMI matches and excluded UMIs with more than 6 Ts on the end facing the oligodT stretch of the primer. Reads were then aligned to the human genome (hg38) using `minimap2`[27] (*-ax splice --secondary=no*).

Cell-type clustering
Both Illumina and R2C2 data were analyzed in the same way independently. First gene expression tables were generated using `featureCounts`[28]. Then these tables were parsed for input into the Seurat R package (v3)[30]. 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, https://github.com/rvolden/Mandalorion-Episode-III).

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

```
-c /path/to/config_file
-m /path/to/NUC.4.4.mat
-s 500
-g /path/to/gencode.v29.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
-O 0,70,0,70
-t 24
-e TGGG,AAAA
```

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

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

Single-cell transcriptomes:

To optimize the alignment step which would consume large amounts of time if the genome would be loaded for each individual cell, i.e. thousands of times, all reads were aligned using `minimap2` (*-ax splice --secondary=no*) and then alignments were split into cells based on read names.

Then Mandalorion was run on the reads and subreads of each individual cell with the modification that the alignment step was skipped and instead `.sam` alignment files were provided. Mandalorion, without read alignment, was run with the following settings:

```
-c /path/to/config_file
-m /path/to/NUC.4.4.mat
-s 500
-g /path/to/gencode.v29.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
-O 0,70,0,70
-t 3
-e TGGG,AAAA
-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
-x /mnt/memorycore1/refs/gmap/hg38
-n
-t 24
-o output_prefix
-d /path/to/output_folder
path/to/gtf_file /path/to/gencode.v29.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`[36] 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[37] using `IgBlast`[35] 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).

Acknowledgements

We thank 10X Genomics for generating full-length cDNA and Illumina sequencing libraries from human PBMCs. We acknowledge funding by the National Human Genome Research Institute/National Institute of Health Training Grant 1T32HG008345-01 (to R.V.), the Hellman Foundation, Santa Cruz Cancer Benefit Group, and National Institute of General Medical Sciences/National Institute of Health Grant 1R35GM133569-01 (to C.V.)

References

1. 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).
2. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
3. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
4. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
5. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
6. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091–1107.e17 (2018).
7. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
8. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
9. Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15**, 563–565 (2018).
10. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
11. Miragaia, R. J. *et al.* Single-Cell Transcriptomics of Regulatory T Cells Reveals Trajectories of Tissue Adaptation. *Immunity* **50**, 493–504.e7 (2019).
12. 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).
13. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter

Droplets. *Cell* **161**, 1202–1214 (2015).

14. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
15. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
16. Tilgner, H. *et al.* Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3* **3**, 387–397 (2013).
17. 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).
18. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv* 459529 (2018) doi:10.1101/459529.
19. 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.
20. 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.* (2018) doi:10.1073/pnas.1806447115.
21. 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).
22. Gupta, I. *et al.* Single-cell isoform RNA sequencing (ScISOr-Seq) across thousands of cells reveals isoforms of cerebellar cell types. *bioRxiv* 364950 (2018) doi:10.1101/364950.
23. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput, error corrected Nanopore single cell transcriptome sequencing. *bioRxiv* (2019).
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. 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).

26. 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).

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

28. 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).

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

30. 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).

31. 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).

32. 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).

33. 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).

34. 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.

35. 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).

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

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

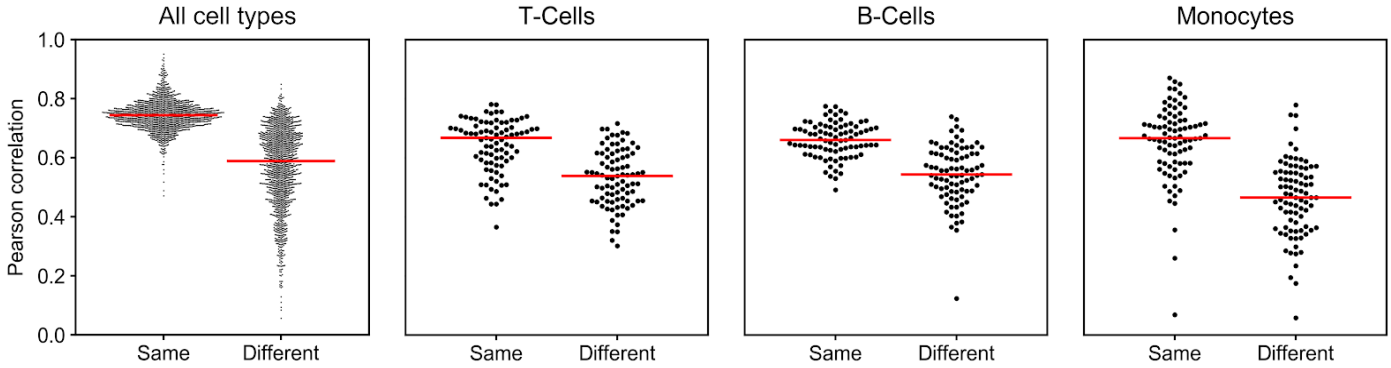
Supplementary Information to

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

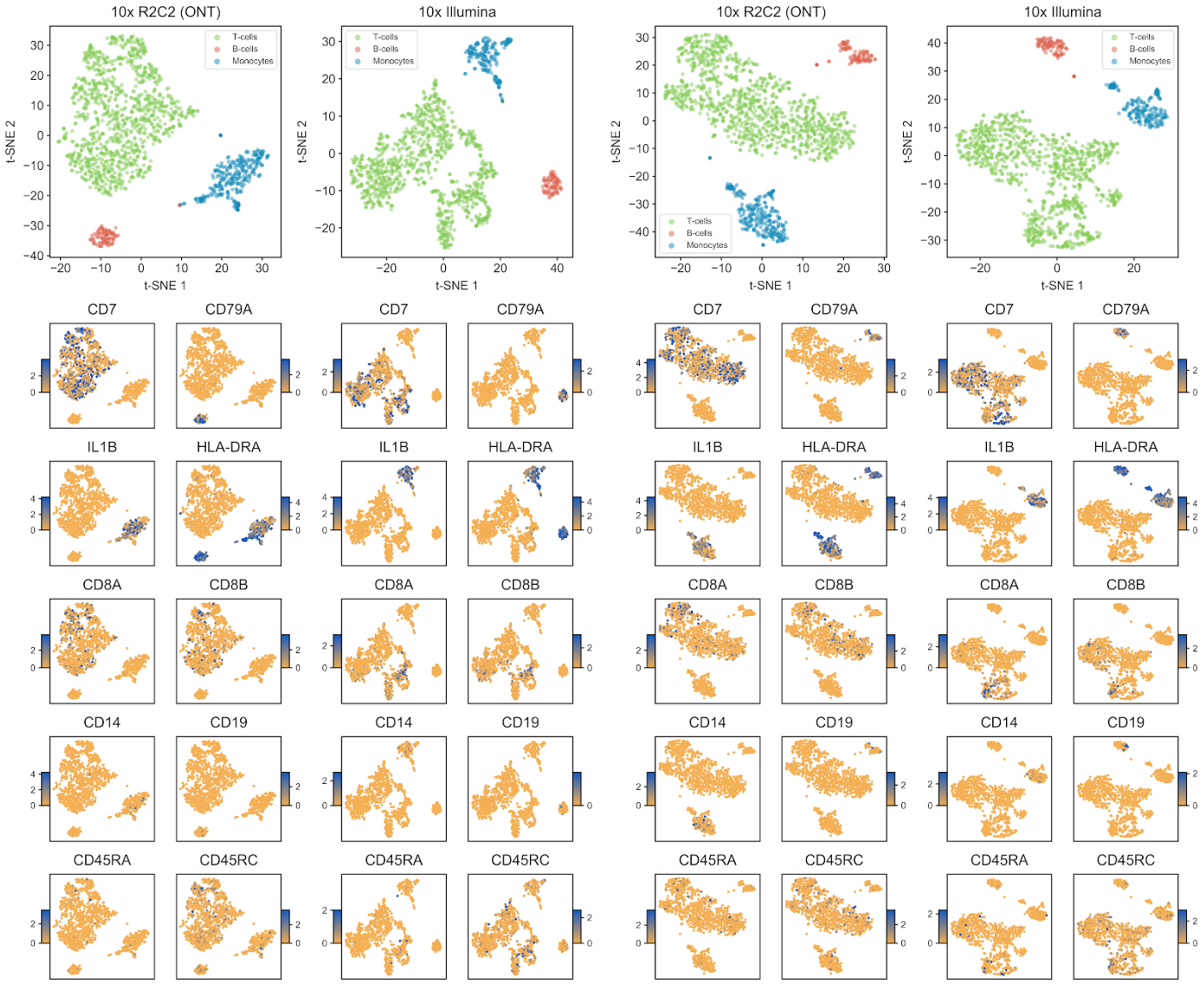
by

Roger Volden and Christopher Vollmers

Table of Contents:



Supplementary Figure S1: Swarm plots of gene expression correlation between R2C2 and Illumina. The median Pearson correlation for each swarm is shown in red. From left to right: (All cell types, same) cells were matched based on their cellular barcode from R2C2 and Illumina. (All cell types, different) R2C2 cells were correlated to a random cell in the Illumina data. The next three swarms were subsampled to 85 points because there are 89 B-Cells. (T-Cells, same) Random T-Cells were correlated between R2C2 and Illumina data. (T-Cells, different) Random R2C2 T-Cells were correlated with random Illumina non-T-Cells. (B-Cells, same) Random B-Cells were correlated between R2C2 and Illumina. (B-Cells, different) Random R2C2 B-Cells were correlated with random Illumina non-B-Cells. (Monocytes, same) Random Monocytes were correlated between R2C2 and Illumina. (Monocytes, different) Random R2C2 Monocytes were correlated with random Illumin non-Monocytes.



Supplementary Figure S2: t-SNE plots with additional marker genes for replicates 1 and 2. As for Figure 2 plots are based on gene expression data as calculated by featureCounts and Seurat. Plots for replicate 1 and replicate 2 are shown on the left and right respectively. Top left: replicate 1 cell type clusters for R2C2 and Illumina. Bottom left: replicate 1 expression heat maps for various marker genes where the two columns on the left are for R2C2 and the right two are Illumina. Top right: replicate 2 cell type clusters for R2C2 and Illumina. Bottom right: replicate 2 expression heat maps for various marker genes where the two columns on the left are for R2C2 and the right two are Illumina. Additional marker genes taken from [14]. The color gradient encodes ln(fold change), where the fold change is comparing that cluster's expression to the rest of the data.

Replicate 1	Basecalled raw reads	Median raw read len	R2C2 Consensus reads
PromethION run 1	14,321,713	3482	5,267,578 (36.78%)
MinION run 1	2,476,608	4460	1,357,914 (54.82%)
PromethION run 2	12,069,611	3913	3,601,214 (29.83%)
MinION run 2	660,000	4019	337,080 (51.07%)
Total	**29,527,932**		**10,563,786 (35.77%)**

Replicate 2	Basecalled raw reads	Median raw read len	R2C2 Consensus reads
PromethION run 1	21,660,888	2711	8,294,332 (38.29%)
MinION run 1	4,865,719	2957	2,291,472 (47.09%)
Total	**26,526,607**		**10,585,804 (39.90%)**

Table S1: Oxford Nanopore Technologies sequencing run and read numbers. Values in parentheses indicate the percentage of raw reads being successfully converted into consensus reads. Note that R2C2 Consensus read numbers indicate consensus reads prior to post-processing. R2C2 Consensus reads after post-processing are given in Table 1.

Replicate 1

R2C2 reads combined into merged read	Number of merged reads	Median accuracy
2	142415	98.3%
3	8290	98.7%
4	876	99.0%
5	133	99.0%
6	33	99.0%
7	13	99.4%
8	4	99.0%
9	2	99.3%
10+	4	99.6%

Replicate 2

R2C2 reads combined into merged read	Number of merged reads	Median accuracy
2	227138	98.0%
3	40953	98.5%
4	11836	98.9%
5	4489	99.0%
6	1677	99.2%
7	834	99.2%
8	441	99.3%
9	204	99.3%
10+	260	99.4%

Table S2: UMIs allow the merging of R2C2 reads originating from the same cDNA molecule.

B cells		Monocytes		T cells		
Gene Symbol	**UIPC**	**Gene Symbol**	**UIPC**	**Gene Symbol**	**UIPC**	
RPS18	0.01	RPS12	0.01	RPS17	0.01	
RPL35	0.01	RPLP2	0.01	RPL35	0.01	
RPS27A	0.01	RPL27	0.01	RPS20	0.01	
RPS23	0.01	RPL35	0.01	RPS10	0.01	
RPS15	0.01	RPS23	0.01	RPL37	0.01	
RPS5	0.01	RPS13	0.02	RPL26	0.01	
RPS13	0.01	RPS20	0.02	RPS23	0.01	
RPS4X	0.02	CST3	0.02	RPL27	0.01	
RPLP2	0.02	S100A4	0.02	RPS25	0.01	
RPS14	0.02	RPL37A	0.02	RPL24	0.01	**Bottom 20**
RPS10	0.02	RPL23	0.02	RPS14	0.01	
RPL18	0.02	RPS14	0.02	RPS29	0.01	
RPL27	0.02	RPLP1	0.02	RPL32	0.01	
FTL	0.03	RPS19	0.02	RPS15A	0.01	
RPS20	0.03	RPL32	0.02	RPS5	0.01	
RPS25	0.03	RPS15	0.02	FAU	0.01	
RPL32	0.03	RPS10	0.02	RPS4X	0.01	
RPS16	0.03	RPS15A	0.02	RPL37A	0.01	
TPT1	0.03	RPL37	0.02	RPLP2	0.01	
RPL34	0.03	S100A8	0.02	RPS18	0.01	
AES	0.49	CEBPB	0.37	IL32	0.28	
HLA-E	0.49	CD68	0.37	GNLY	0.3	
GNB2	0.5	EIF4A1	0.37	IDS	0.3	
CRIP1	0.5	FCGRT	0.37	SEPT9	0.3	
ARPC1B	0.52	HLA-E	0.4	LINC-PINT	0.3	
TRAF4	0.6	UPP1	0.42	DPP7	0.31	
DPP7	0.61	KLF4	0.43	ITGB2	0.31	
DDX5	0.63	CARD19	0.46	UBE2D3	0.31	
ADGRE5	0.68	PSAP	0.47	DDX5	0.32	
EIF4A1	0.68	KLF2	0.47	ATG2A	0.38	**Top 20**
CD83	0.69	VASP	0.49	IRF1	0.48	
FAM129C	0.71	PABPC1	0.5	PIK3IP1	0.53	
NFKBID	0.74	ITGB2	0.53	SPOCK2	0.54	
IRF1	0.76	ITGAX	0.55	PLK3	0.59	
TAPBP	0.86	GRN	0.56	ADGRE5	0.61	
NR4A2	0.88	LST1	0.77	NR4A2	0.62	
TBC1D10C	0.89	TRABD	0.79	POLR2A	0.63	
EZR	0.95	TOM1	0.81	CD6	0.77	
LMNA	0.97	SLC11A1	0.99	TMC8	0.78	
IGHM	1	LMNA	1.21	LMNA	0.8	

Table S3: Isoform diversity in B cells, T cells, and Monocytes. The 20 genes with the most or least isoform diversity as determined by the Unique Isoforms per Cell (UIPC) measure are shown for the indicated cell-types.