Supplemental Materials and Methods, Figures, and Tables

to

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

by Volden et al.

www.pnas.org/cgi/doi/10.1073/pnas.1806447115

Supplemental Material and Methods

R2C2 library preparation

SPRI Bead DNA extractions

After Rolling Circle Amplification, 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.

Data Analysis

After sequencing the HMW DNA on the ONT MinION using either ligation (LSK108) or transposase (RAD004) based kits, the resulting raw data was basecalled using the albacore (version 2.1.3) read_fast5_basecaller script with the following settings:

Ligation 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(14) 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*)

C3POa.py is a wrapper script that performs the following steps:

1) Tandem repeats in each raw read are detected using a modified EMBOSS WATER(21–23) 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. 1).

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(15) 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 sequence is used as consensus in the following steps.

3.) Complete and incomplete subreads are aligned to the consensus sequence using minimap2(24) 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(16) algorithm which error-corrects the consensus using the following command:

```
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
```

To speed up consensus calling, we divided raw reads into bins of 4000 and used GNU Parallel(25) to run multiple instances of C3POa.py .

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 (26) 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 defineAndQuantifyWrapper.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 -f
/path/to/config_file
```

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
```

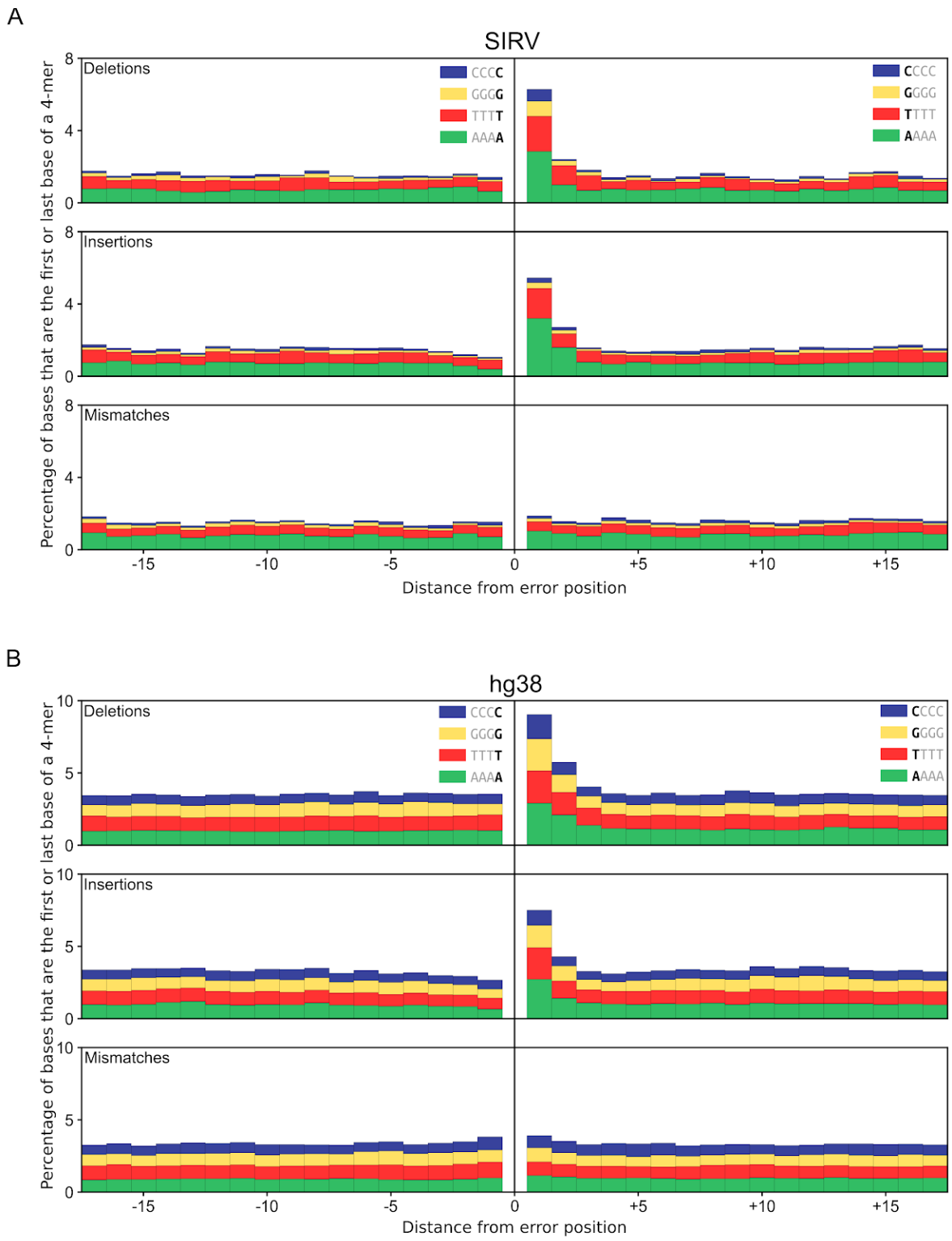

Fig. S1 Sequence context of error in R2C2 reads.
The occurrence of 4-mers (CCCC,GGGG,TTTT, or AAAA) around R2C2 sequencing deletions, insertions, and mismatches is shown as stacked bar plots for R2C2 reads covering SIRV cDNA (A) or human B cell cDNA (aligned to hg38) (B).

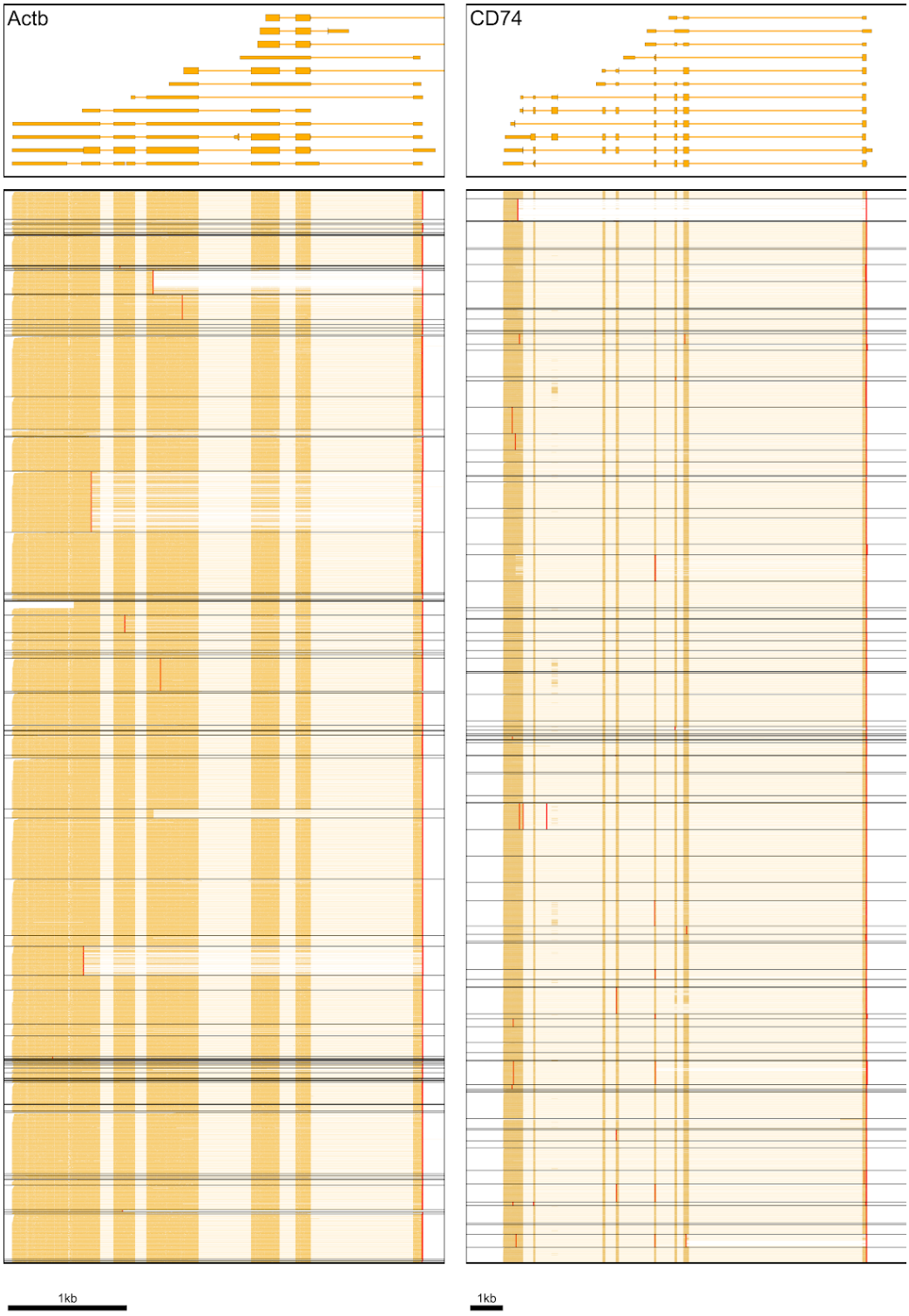

Fig. S2: R2C2 consensus reads of single cell cDNA. *Genome browser view of Transcriptome annotation and up to 200 R2C2 consensus reads per cell is shown of the indicated human gene loci. R2C2 consensus reads from different cells are separated by black lines. Transcript and read direction is shown by colors (Blue: + strand , Yellow: - strand). Red lines indicated TSSs identified based on the same cDNA using Tn5Prime based Illumina sequencing.*

Fig. S3: Predicted amino acid sequences of CD19 isoforms lack FMC63 CAR-T cell epitope

CD19 isoforms identified by Mandalorion based on single cell R2C2 consensus reads were manually aligned to the CD19 protein reference (on top). Red letters indicate mismatches to the CD19 protein reference. The FMC63 epitope is shown in green.

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 S1: R2C2 run statistics

Sequencing Platform	Protocol	Number of adapter-to-adapter cDNA molecules sequenced for $2500	Median Accuracy
Illumina HiSeq4000[1]	Standard RNAseq	N/A	>99%
Illumina HiSeq4000[1]	spISO-seq	N/A[4]	>99%
Illumina HiSeq4000[1]	SLR	2,000[5] (1% of all assembled reads)	~98.5
PacBio Sequel[2]	Iso-Seq	300,000 (50% of all ccs reads)	~98%
ONT MinION[3]	1D cDNA	750,000-1,250,000[6]	~86%
ONT MinION[3]	1D2 cDNA	150,000-250,000[7]	~95%
ONT MinION[3]	R2C2	1,000,000-1,600,000[8]	~94%

Table S2: Sequencing technology comparison: *Approximate numbers are based on public data (PacBio, SLR(2)) and data produced in our lab (PacBio, ONT). Data from multiplexed experiments was extrapolated to $2500 worth of sequencing runs. Accuracy is calculated from minimap2 alignments using the following formula (matches/(matches+mismatches+indels)).*

1) Illumina cost was based on a single lane of a PE150 run at ~$2500 (UC Berkeley UC-internal rate)

2) PacBio cost was calculated based on two $1300 SMRT cells (UC Berkeley UC-internal rate) producing 150,000 adapter-to-adapter reads each.

3) ONT cost was calculated based on three flowcells at $790 (cost when purchasing 12 flowcells) or five flowcells at $500 (cost when purchasing 48 flowcells). Read output per flowcell is dependent on sample quality, sequencing protocol used, and flowcell condition. All ONT read numbers are therefore based on data produced in our lab and may vary in other laboratories.

4) spiSO-seq does not assemble molecules, it instead generates read cloud containing long-distance information.

5) Because SLR does not assemble its 3' adapter, which lies past the polyA tail, we defined a SLR read as "adapter-to-adapter" if it aligned within 20bp of both ends of synthetic spike-in molecules.

6) A single ONT flowcell produces 1,000,000 1D cDNA reads. Only 25% of these are adapter-to-adapter.

7) A single ONT flowcell produces 50,000 1D2 cDNA reads most of which are adapter-to-adapter.

8) A single ONT flowcell produces 316,000 R2C2 all of which are adapter-to-adapter.

cDNA Generation

Reverse Transcription

```
>Oligo-dT-smartseq2
/5Me-isodC/AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

>TSO1_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  rUGA  ArU  rUC  TGGTrGrGrG
>TSO2_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  ACrU  CrU  GrU  TGGTrGrGrG
>TSO3_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  CrUC  rUG  rUA  TGGTrGrGrG
>TSO4_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  rUAG  rUA  CrU  TGGTrGrGrG
>TSO5_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  GGrU  CrU  rUG  TGGTrGrGrG
>TSO6_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  ArUA  GrU  ArU  TGGTrGrGrG
>TSO7_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  rUCC  rUA  rUC  TGGTrGrGrG
>TSO8_Nextera
TCGTCGGCAGCGTCAGATGTGTATAAGAGArCAG  CArU  rUC  GrU  TGGTrGrGrG
```

cDNA Amplification

```
>ISPCR
AAGCAGTGGTATCAACGCAGAGT
>Nextera_Primer_A
AATGATACGGCGACCACCGAGATCTACAC  [8bp  i5  index]  TCGTCGGCAGCGTCAGATG
```

Splint Generation

```
>Lambda_F_ISPCR
ACTCTGCGTTGATACCACTGCTT AAAGGGATATTTTCGATCGCTTG
>Lambda_R_NextA
ATCTCGGTGGTCGCCGTATCATT TGAGGCTGATGAGTTCCATATTTG
```

Table S3 Oligos used in the manuscript

All oligos are shown 5'->3' and were ordered from Integrated DNA Technologies (IDT). Lower case 'r' indicates RNA bases. Spaces in sequences are for visual emphasis only.