UNIVERSITY OF CALIFORNIA
SANTA CRUZ

UTILIZING THE R2C2 LONG READ SEQUENCING PROTOCOL FOR VARIOUS APPLICATIONS

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Kayla D. Schimke

September 2024

The Dissertation of Kayla D. Schimke is approved:

Professor Christopher Vollmers, Chair

Professor Angela Brooks

Professor Manuel Ares

Professor Joshua Arribere

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Utilizing the R2C2 Long Read Sequencing Protocol for Various Applications

Kayla D. Schimke

Abstract

The Rolling Circle Amplification to Concatemeric Consensus (R2C2) protocol was developed to increase the accuracy of Oxford Nanopore Technologies (ONT) long reads for transcriptomic analysis. Originally designed to transform cDNA input into long reads containing multiple copies of individual molecules, this protocol has a wide range of applications using various types of input. I demonstrate the versatility of R2C2 with three general types of input: Illumina libraries, 10X single cell RNA libraries, and plasmids. I developed the PLNK (Processing Live Sequencing experiments) software to monitor ONT MinION runs of Illumina R2C2 libraries to analyze the read quality, library content and coverage over target regions. I used 10X single cell RNA libraries as input for R2C2 and analyzed them with Mandalorion to produce cell type specific transcriptomes. Finally, I shortened the R2C2 protocol to allow for plasmid input and created Chopper, a software pipeline which produces accurate assemblies of full length plasmids.

Acknowledgements

The text of this dissertation includes reprints of the following previously published material: Sequencing Illumina libraries at high accuracy on the ONT MinION using R2C2 (Zee et al. 2022) and Sequencing full-length plasmids with ONT long-reads using R2C2 (Schimke and Vollmers 2024). 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 pursue my goal of becoming a doctor. I know I had planned to be a veterinarian when I was younger but a Doctor of Philosophy deals with fewer people. My family are some of my favorite people and I wouldn't have made it as far as I have without their love and encouragement. I would also like to thank my lab mates Dori Deng and Alexander Zee; we started graduate school together and we ended it together. Dori is a force to be reckoned with both in personality and in science and I strive to be half the scientist that she is. Alex brings vitality to life and science, and I will always remember his joy. I look forward to seeing what they do in the future.

Finally, I would like to thank Chris Vollmers for being an outstanding mentor and supporting my professional goals. I didn't think I would be in Chris' lab at the beginning of my graduate career but there truly could not have been a better place for me. He pushed me to expand my computational and molecular biology skills and helped me build my teaching skills. He is the kind of mentor I strive to be, and I will miss working with him.

Utilizing the R2C2 Long Read Sequencing Protocol for Various Applications

An overview of the R2C2 protocol

Every being is defined by its DNA but uncovering the genetic variations between species, and even individuals of the same species, hinges on reliable sequencing techniques. After sequencing the human genome for the first time, genomics researchers pushed toward making sequencing faster, cheaper and more accurate, eventually creating a short read sequencing platform that became the standard.

Short read sequencing, championed by Illumina, involves fragmenting and amplifying input DNA/cDNA to maintain accuracy and ensure even coverage. While the quality of Illumina short read sequencing technology is indisputable, it has a few shortcomings. For whole genome analysis, the limited read length prevents accurate assemblies within repetitive regions of the genome. For transcriptomic analysis, short reads fail on two accounts: declining coverage towards the ends of transcripts and missing portions within transcripts caused by read fragmentation. Both of these issues prevent the assembly of full length transcripts and limit isoform analysis.

Long read sequencing techniques, such as those offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), are better suited to transcriptomic analysis and aid whole genome analysis. Their extended read length ensures sequencing of full length RNA transcripts while maintaining coverage. A longer read length also helps capture large repeats, as a single read is more likely to span the entire region. Unfortunately, these long read technologies can’t match the read accuracy that Illumina short reads exhibit. Many studies use a combination of long and short read sequencing to improve the contiguity and accuracy of their data,

while others have developed protocols or tools for improving the quality of long reads alone.

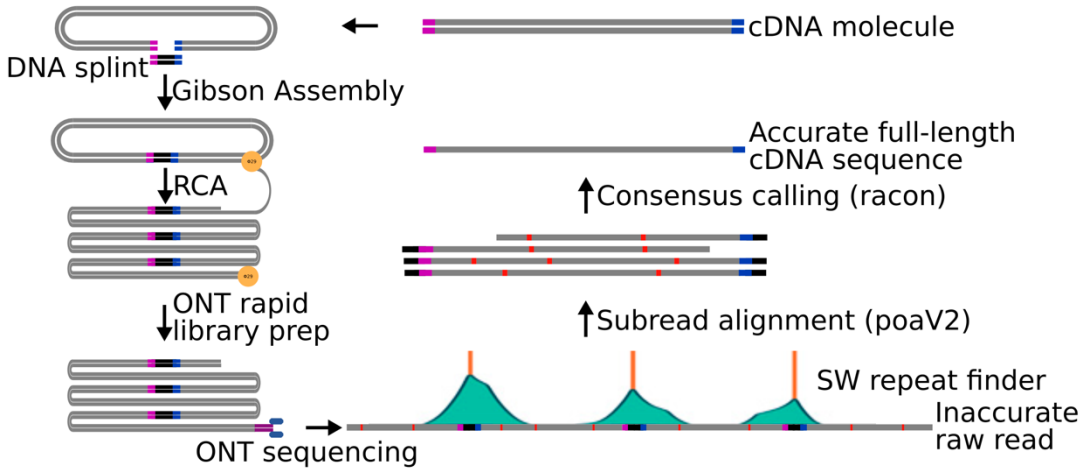

Figure 1. R2C2 method overview. cDNA is circularized using Gibson Assembly, amplified using RCA, and sequenced using the ONT MinION. The resulting raw reads are split into subreads containing full-length or partial cDNA sequences, which are combined into an accurate consensus sequence using our C3POa workflow, which relies on a custom algorithm to detect DNA splints as well as poaV2 and racon. (Volden et al. 2018)

The Vollmers lab developed the Rolling Circle to Concatemeric Consensus (R2C2) protocol to generate high quality, full-length transcriptomic sequencing data from ONT sequencers (Volden et al. 2018) (Fig. 1). After generating cDNA from a RNA sample, the molecules are circularized through Gibson assembly with an known DNA splint before rolling circle amplification (RCA) with Phi-29 polymerase. The process produces long strands of DNA containing multiple copies of the original cDNA molecules with the splint sequence between each copy. The sample can then be prepared for sequencing on an ONT instrument and basecalled as normal.

Given the unique structure of R2C2 molecules, the Vollmers lab developed a computational tool called C3POa (Concatemeric Consensus Caller with Partial Order

alignments) which aligns the DNA splint sequence to each read to find the repeats of the original molecule and dividing the raw read into subreads. The subreads are aligned to each other to create a consensus sequence. The combination of R2C2 and C3POa analysis increased the then ~92% accurate ONT reads to >99% and continues to push that limit as the technology improves.

While originally developed to improve the quality of long read data for transcriptomics, the Vollmers lab has used the R2C2 protocol for multiple applications, three of which will be discussed in this thesis. One application of R2C2 is to amplify Illumina libraries for ONT sequencing in order to assess the quality and shorten the processing time of the library. I developed the PLNK (Processing Live Nanopore experiments) software which runs alongside an ONT MinION sequencing run, capturing individual FAST5 files to analyze the input Illumina libraries. PLNK successfully keeps up with MinION runs while basecalling, consensus calling and demultiplexing the reads in individual FAST5 files before aligning them to a reference and analyzing the coverage and percentage of reads that fall within targeted regions.

I also applied R2C2 to sequence 10X single cell RNA libraries. The Vollmers lab previously demonstrated the effectiveness of sequencing 10X libraries with R2C2 in a Genome Biology paper focused on human immune cells (Volden and Vollmers 2022). I applied this protocol to sequence 10X libraries from lung tissue provided by the Tabula Sapiens Consortium. The R2C2 10X data was then processed using C3POa for consensus calling, Cell Ranger for demultiplexing and Mandalorion (https://github.com/christopher-vollmers/Mandalorion) for isoform analysis to produce cell type specific transcriptomes.

Finally, I modified the R2C2 protocol to amplify plasmids for sequencing on an ONT P2 Solo. The initial step of the R2C2 protocol is circularizing cDNA via Gibson assembly. Plasmids are naturally circles, making them the perfect input for RCA and allowing us to skip the first step of the R2C2 protocol. Using the modified version of R2C2 to amplify and sequence plasmids increased the accuracy of the raw reads and ensured a high quality, full length assembly.

Processing live nanopore experiments with PLNK

This section is adapted from **Sequencing Illumina libraries at high accuracy on the ONT MinION using R2C2** (Zee et al. 2022)

Sequencing Illumina libraries at high accuracy on the ONT MinION using R2C2

Alexander Zee[1,4], Dori Z.Q. Deng[2,4], Matthew Adams[2,4], Kayla D. Schimke[1,4], Russell Corbett-Detig[1], Shelbi L. Russell[2], Xuan Zhang[3], Robert J. Schmitz[3], and Christopher Vollmers[1,*]
1) Department of Biomolecular Engineering,
2) Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, California 95064, USA;
3) Department of Genetics, University of Georgia, Athens, Georgia 30602, USA
4) These authors contributed equally to this work.
*) Corresponding author. Email: vollmers@ucsc.edu

Abstract

High-throughput short-read sequencing has taken on a central role in research and diagnostics. Hundreds of different assays take advantage of Illumina short-read sequencers, the predominant short-read sequencing technology available today. Although other short-read sequencing technologies exist, the ubiquity of Illumina sequencers in sequencing core facilities and the high capital costs of these technologies have limited their adoption. Among a new generation of sequencing technologies, Oxford Nanopore Technologies (ONT) holds a unique position because the ONT MinION, an error-prone longread sequencer, is associated with little to no capital cost. Here we show that we can make short-read Illumina libraries compatible with the ONT MinION by using the rolling circle to concatemeric consensus (R2C2) method to circularize and amplify the short library molecules. This results in longer DNA molecules containing tandem repeats of the original short

library molecules. This longer DNA is ideally suited for the ONT MinION, and after sequencing, the tandem repeats in the resulting raw reads can be converted into high-accuracy consensus reads with similar error rates to that of the Illumina MiSeq. We highlight this capability by producing and benchmarking RNA-seq, ChIP-seq, and regular and target-enriched Tn5 libraries. We also explore the use of this approach for rapid evaluation of sequencing library metrics by implementing a real-time analysis workflow.

Introduction

Over the past 15 years, high-throughput short-read sequencing technology has revolutionized biological, biomedical, and clinical research. Hundreds of sequencing-based methods exist today to query gene expression (RNA-seq) (Mortazavi et al. 2008), chromatin state (chromatin immunoprecipitation [ChIP]-seq and ATAC-seq) (Barski et al. 2007; Buenrostro et al. 2013), protein abundance (Stoeckius et al. 2017) and, of course, to aid the assembly of genomes (Burton et al. 2013)—among many other applications. All of these methods produce a final sequencing library that contains ~200- to 600-bp double-stranded DNA molecules with ends of a known sequence. In the vast majority of cases, these ends are Illumina sequencing adapters.

Despite the existence of other sequencing technologies, Illumina has been the dominating short-read sequencing technology over the past decade. However, because of the high capital cost of Illumina short-read instruments, all but the most well equipped laboratories outsource their Illumina sequencing to core facilities. Although this provides access to the most recent sequencing technology, this

outsourcing can lead to long delays between running an experiment and receiving results. Therefore, placing a benchtop sequencer with capabilities comparable to an Illumina sequencer in most molecular biology and diagnostic laboratories could be truly transformative by accelerating as well as fully integrating genomics assays into standard laboratory workflows. In a molecular biology laboratory, it would speed up developing or establishing new types of sequencing libraries. In a diagnostic laboratory, it could enable fast sample turn-around as well as encourage the transition away from diagnostic methods like fluorescence in situ hybridization (FISH), which is still routinely used for the detection of gene fusions in certain cancers despite having a >20% false-negative rate and the availability of more accurate sequencing-based replacements (Ali et al. 2016; Nohr et al. 2019).

Over the past few years Oxford Nanopore Technologies (ONT) sequencers have rapidly matured. Currently, the ONT MinION sequencer's base throughput (up to 30 Gb per flow cell) can exceed that of the Illumina MiSeq sequencer (18 Gb for a 2 x 300-bp run). Additionally, this throughput comes with tunable read length, so a successful MinION run can in theory produce 10 million 3-kb reads or 5 million 6-kb reads. Further, the MinION sequencer is only a fraction of the cost of other high-throughput sequencers. However, standard per-base sequencing accuracy of the newest base-calling software guppy5 is only ~96% and is dominated by insertion and deletion errors, which are almost absent in Illumina data. Furthermore, ONT MinION's sequencing accuracy declines with shorter reads (Thirunavukarasu et al. 2021).

Here, we implemented a simple workflow that converts almost any Illumina sequencing library into DNA of lengths optimal for the ONT MinION and generates

data at similar cost and accuracy as the Illumina MiSeq. We made this possible by using the previously published and optimized rolling circle to concatemeric consensus (R2C2) method (Volden et al. 2018; Byrne et al. 2019; Adams et al. 2020; Cole et al. 2020; Vollmers et al. 2021; Volden and Vollmers 2022). R2C2 circularizes dsDNA libraries and amplifies those circles using rolling circle amplification to create long molecules with multiple tandem repeats of the original molecule's sequence. These long molecules can then be sequenced on ONT instruments to generate long raw reads, which are then computationally processed into accurate consensus reads. In previous studies focused on full-length cDNA molecules, we have achieved median read accuracies of 99.5% with this method (Vollmers et al. 2021). Because Illumina libraries are shorter than full-length cDNA, we modified the R2C2 protocol to generate a large number of shorter MinION raw reads while maintaining consensus accuracy levels on par with the Illumina MiSeq sequencer.

We benchmark this extension of the R2C2 method by converting and sequencing RNA-seq and ChIP-seq, as well as regular and target-enriched genomic DNA Tn5 Illumina libraries. We implemented a computational workflow for demultiplexing Illumina library indexes from R2C2 data and have, where possible, relied on established analysis workflows for downstream analysis originally developed for Illumina data. If R2C2 and Illumina data required different computational approaches, that is, assembly and variant calling, we chose the optimal tool for either data type.

To take advantage of the real-time data generation of ONT sequencers, we also developed Processing Live Nanopore Experiments (PLNK), for monitoring and rapid evaluation of sequencing runs. PLNK uses several tools to base-call,

demultiplex, and map reads as they are generated. PLNK then reports, in real-time, run features like what percentages of reads belong to each library in a library pool, what percentages of reads in each library map to a list of target regions, and what the read coverage of these target regions is for each library.

This work was performed with the aim of evaluating whether R2C2 and the associated computational methods C3POa and PLNK could be used to replace and potentially even improve on dedicated Illumina sequencers for the analysis of short-read libraries.

Results

To generate R2C2 data for a diverse selection of Illumina libraries, we processed and sequenced (1) Illumina RNA-seq libraries of the human A549 cancer cell line, (2) Illumina ChIP-seq and input libraries of soybean samples, (3) Illumina Tn5-based genomic DNA libraries of a Wolbachia-containing Drosophila melanogaster cell line, and (4) Illumina Tn5-based genomic DNA libraries generated from the lung cancer cell lines NCI-H1650 and NCI-H1975, which we enriched for the protein-coding regions of approximately 100 cancer-relevant genes (Fig. 2).

To convert these Illumina libraries into R2C2 libraries, we circularized them using Gibson assembly (NEBuilder/NEB) with DNA splints compatible with Illumina p5 and p7 sequences. After the DNA circles are amplified with rolling circle amplification using Phi29 polymerase, we fragmented and size-selected the resulting high-molecular-weight DNA. We then sequenced this DNA on the ONT MinION using the LSK-110 ligation chemistry and 9.4.1 flow cells. We generated between 4 and 9.5 million raw reads per MinION flow cell. All data were then base-called with the

guppy5 *dna_r9.4.1_450bps_sup.cfg* model and consensus called using C3POa (v.2.2.3; https://github.com/rvolden/C3POa).

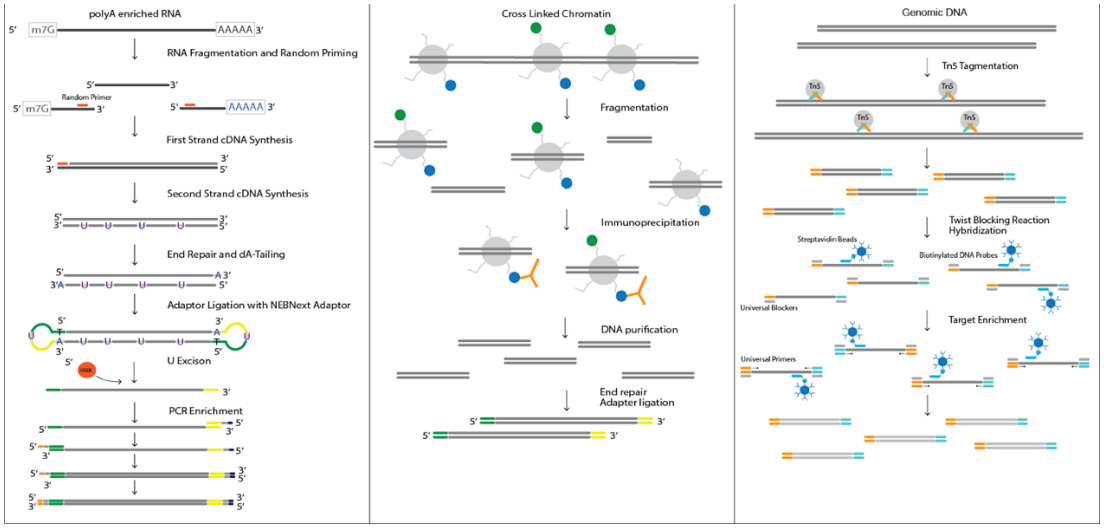

Figure 2. Illumina library preparation overview. Illumina RNA-seq, ChIP-seq, and Tn5-based genomic libraries (regular and enriched) were generated from different samples. The Illumina libraries were then circularized and amplified using rolling circle amplification (RCA). The resulting DNA, containing tandem repeats of Illumina library molecules, was then prepared for sequencing on the ONT MinION sequencer and monitored with PLNK.

Real-time analysis of Illumina library metrics using PLNK

To enable the real-time monitoring of sequencing runs and the rapid evaluation of metrics of libraries sequenced in those runs, we created the computational pipeline Processing Live Nanopore Experiments (PLNK). PLNK controls real-time basecalling, raw read processing into R2C2 consensus reads, demultiplexing of R2C2 reads, and the alignment of demultiplexed R2C2 reads to a genome. Based on the resulting alignments and the user-defined regions of interest, PLNK then determines the on target percentage and resulting target coverage for each demultiplexed sample. PLNK runs alongside a MinION sequencing run, tracking the

creation of new FAST5 files and processing them individually in the order they are generated. To do this, PLNK controls several external tools: guppy5 for basecalling, C3POa for R2C2 consensus generation, C3POa post-processing for demultiplexing (based on splint sequences and Illumina indexes), and mappy (minimap2 Python library) for aligning reads to a provided genome (Fig. 3A).

To test whether our pipeline could keep up with ONT MinION data generation and provide real-time analysis, we simulated ONT MinION runs using FAST5 files from three previously completed sequencing experiments involving target-enriched Tn5, ChIP-seq, and RNA-seq data (Fig. 2). We used the FAST5 files' metadata to determine the time intervals at which files were generated by the MinKnow software and copied the FAST5 files to a new output directory at those intervals. We then started PLNK to monitor the generation and control the processing of FAST5 files in this new output directory.

First, we simulated the real-time analysis of the target-enriched Tn5 data. Using a desktop computer and limiting PLNK to the use of eight CPU threads and two Nvidia RTX2070 GPUs, the pipeline processed sequencing data at the same rate that a single MinION produced FAST5 files. Importantly, both the library composition (percentage of demultiplexed reads assigned to either sample; NCI-H1650 and NCI-1975), as well as the percentage of reads on target, stabilized after less than an hour and agreed very well with the numbers generated from the whole dataset (Fig. 3B). Additionally, throughout the run, PLNK reported the overall coverage of target regions in real-time.

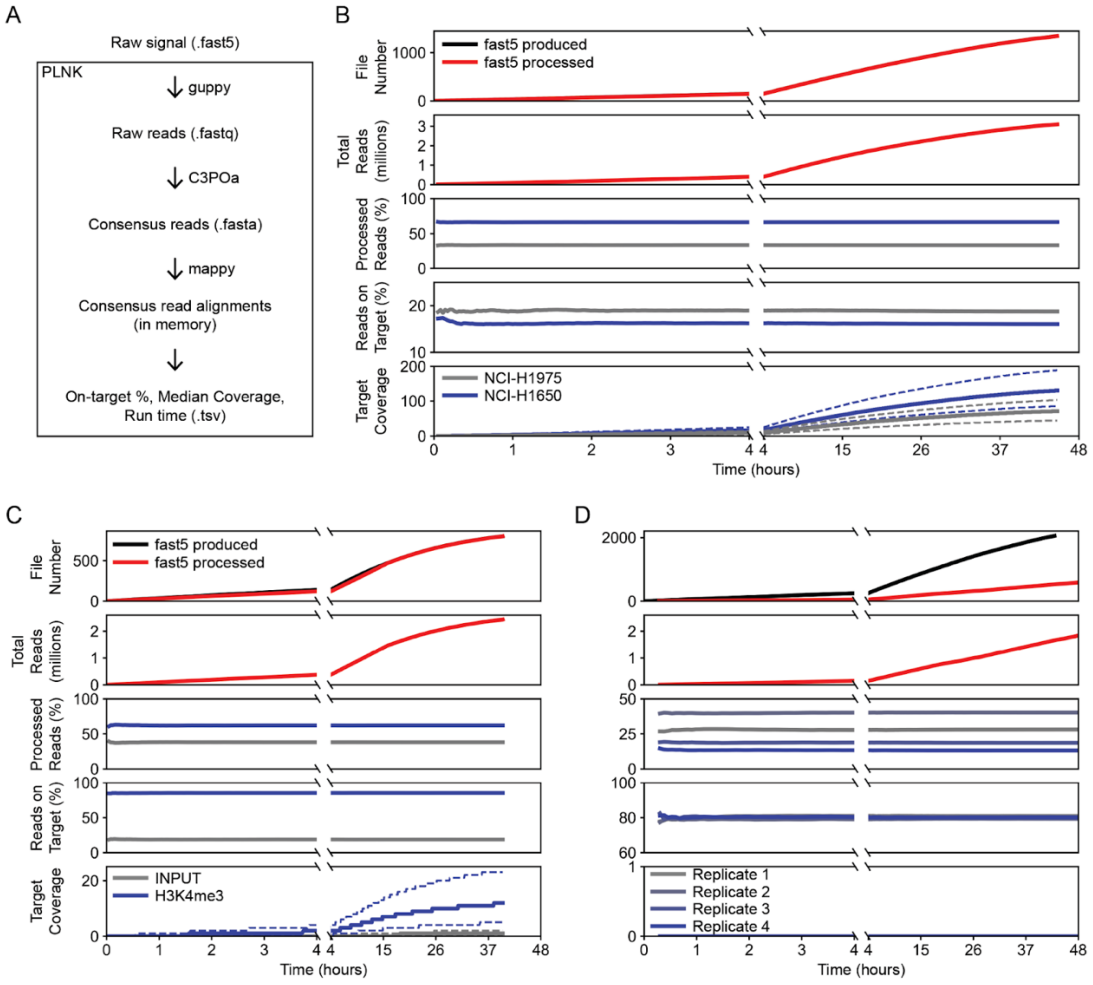

Figure 3. Real-time characterization of Illumina sequencing libraries. (A) Diagram of PLNK functionality; FAST5 files processed in the order they are produced. PLNK controls guppy5 for basecalling, C3POa for consensus calling, and mappy for alignment, as well as calculates metrics based on those alignments. (B–D) Simulation of real-time analysis for enriched Tn5 (B), ChIP-seq (C), and RNA-seq (D) libraries. For each time point, panels from top to bottom show (1) the number of FAST5 files that are produced and processed, (2) the number of demultiplexed reads produced by guppy5/C3POa/demultiplexing, (3) the percentage of reads associated with each library in the sequenced pool, (4) the percentage of reads overlapping with target regions, and (5) the median read coverage of bases in the target regions.

When we simulated the analysis of ChIP-seq and RNA-seq experiments, PLNK kept up with ChIP-seq but not with the RNA-seq experiment (Fig. 3CD).

Because the RNA-seq experiment produced the largest amount of data of the three experiments, this was not unexpected. In both cases, however, library composition and on-target percent both stabilized within the first hour of sequencing and reflected the number derived from the complete dataset. This means that the library composition and quality of target-enriched Tn5 libraries (as measured by reads overlapping target areas), ChIP-seq libraries (as measured by reads overlapping with peak areas, promoters, or gene bodies—depending on targeted histone mark), and RNA-seq libraries (as measured by reads overlapping with exons) can be determined with minimal sequencing time.

The bottleneck for analysis in our desktop computer setup seemed to be the guppy5 based basecalling using the slower yet most accurate “sup” base-calling configuration. Although we could use a faster, less accurate setting to keep up with even the fastest data-producing experiments, using the most accurate model means the data can be used for in-depth analysis once the run has completed and PLNK has processed all the files, without the need to re-basecall the raw data.

Overall, this suggests that PLNK can be used to monitor ONT MinION sequencing runs in real-time. Monitoring the PLNK’s output makes it possible to stop ONT sequencing runs when the goal of an experiment is achieved. For the rapid evaluation of library pools, this could be 1 hour into a run once library composition and quality metrics have stabilized. For run monitoring, this could be several hours into a run once a specific coverage of defined target regions is reached. In both cases, a run can be stopped, allowing the ONT MinION flow cell to be flushed, stored, and ultimately reused.

Discussion

The capabilities of the dominant Illumina sequencing technology—producing massive numbers of short reads—have shaped the development of sequencing-based assays more than any other single factor.

Although long-read sequencers by Pacific Biosciences (PacBio) and ONT have now superseded Illumina instruments as the gold standard technology for genome assembly, producing libraries for these long-read sequencers requires relatively large amounts of high-quality DNA material. In many cases, both DNA input amount and/or quality of a sample may not match these requirements, leaving amplification-based short-read sequencing as the only option to extract large amounts of sequencing data from that sample.

Beyond the sequencing and assembly of genomes, there are hundreds of assays adapted for short reads. These assays are highly diverse and require different levels of read numbers and accuracy, and many, like standard RNA-seq, ChIP-seq, or targeted sequencing of PCR amplified genomic DNA, are unlikely to ever take advantage of the raw read length ONT and PacBio sequencers provide. However, there have been several studies to take advantage of long-read sequencing instruments in sequencing shorter molecules. Some assays (OCEAN, MAS-Iso-Seq) work by either concatenating (Al'Khafaji et al. 2021; Thirunavukarasu et al. 2021) or otherwise preparing (Baslan et al. 2021) short molecules for sequencing on the PacBio or ONT instrument. Although these assays can generate more short reads, they have to contend either with the high cost of the PacBio Sequel IIe sequencer or with the low per-base accuracy of raw ONT reads, which even with the latest guppy5 algorithm is only 96% in our hands. Even at 96%, this

ONT raw accuracy is likely sufficient for certain applications like ChIP-seq, where reads simply have to be aligned to a genome and counted. For these applications, preparing and sequencing short-read libraries directly on an ONT sequencer is a straightforward option. This approach would also allow the usage of native ONT barcoding strategies, which are more robust at low accuracy. However, sequencing short-read libraries directly on ONT sequencers has the downside that these sequencers have reduced output when sequencing short molecules <1 kb. Specifically, when sequencing molecules that are ~500 nucleotides in length, the overall base-output of an ONT MinION flow cells seems to vary between 3 Gb (Glinos et al. 2022), 4 Gb (this study), and 9 Gb (Pardo-Palacios et al. 2021)—far below the 30 Gb maximum output these flow cells can achieve when sequencing longer molecules. There is therefore room to optimize ONT library preparations for short-read sequencing.

Taking inspiration from the highly accurate but throughputlimited PacBio Iso-Seq and HiFi workflows, circularizing-based (R2C2, INC-seq, and HiFRe) (Li et al. 2016; Volden et al. 2018; Wilson et al. 2019) methods have been developed to trade throughput for accuracy on ONT MinION and PromethION sequencers. Using a modified R2C2 method we present here, we show that we can convert any Illumina sequencing library with double-stranded adapters—PCR-free “crocodile adapter”–style libraries will not work—into an R2C2 library that is several kilobases long and therefore takes full advantage of the ONT MinION’s throughput. The close to optimal base-output of up to 24 Gb (9.5 million raw reads Å~ 2.5-kb average read length) when sequencing R2C2 libraries allowed us to produce not only more accurate reads but also a higher number of total reads than regular ONT 1D libraries of the same

short-insert Illumina libraries. In fact, the throughput and accuracy of R2C2 were comparable to Illumina MiSeq 2 Å~ 300-bp runs.

By generating up to 8.99 million reads (8.1 million demultiplexed) with a per-base accuracy of 98.87% (Illumina MiSeq read 1: 99.47%; read 2: 98.57%) from a single ONT MinION flow cell, this approach can compete with the Illumina MiSeq and other benchtop Illumina sequencers on accuracy and cost—even without taking instrument cost into account. Improved consensus tools (Silvestre-Ryan and Holmes 2021), the consistently improving ONT sequencing chemistry and basecallers, and the imminent release of a much cheaper ONT PromethION variant (P2Solo) all have the potential to further skew both accuracy and throughput comparison in R2C2's favor in the near future. Not only might improving ONT sequencing chemistry improve throughput, but it might also mitigate the considerable variability in throughput we see in R2C2 read output (4 to 9 million reads).

We have shown the capabilities and limitations of this approach here by evaluating the conversion of RNA-seq, ChIP-seq, genomic Tn5, and target-enriched genomic Tn5 libraries. The R2C2 data were more than accurate enough to demultiplex Illumina libraries based on their i5 and i7 indexes. Furthermore, RNA-seq data produced with R2C2 were almost entirely interchangeable with data produced by the Illumina MiSeq. Library metrics derived from R2C2 data generated from ChIP-seq and target-enriched Tn5 libraries showed library metrics very similar to those determined from data generated by Illumina sequencers. One notable exception to this were insert length distributions of Illumina libraries, where R2C2 produced longer insert distributions than Illumina sequencers, which are known to prefer shorter molecules enough to affect analysis outcomes (Gohl et al. 2019). For the ChIP-seq

experiment, but no other experiment in this paper, R2C2 reads also had a slightly higher GC content, which made the Illumina/R2C2 comparison less quantitative than it was, for example, in the RNA-seq experiment. For germline variant calling, R2C2 reads analyzed with PEPPERMargin- DeepVariant produced variant calls highly similar to Illumina/DeepVariant variant calls, with no false positives (precision 100%) and only three false negatives (recall 97.4%), two of which were next to homopolymers, which are known to be a challenge for ONT sequencers.

Taken together, we have established that R2C2 can be used as a drop-in replacement for many sequencing-based applications that would usually demand a dedicated short-read Illumina sequencer. One important thing to note is that R2C2 adds complexity to an ONT experiment. Sequencing a short-read library using regular 1D ONT sequencing at most requires the library to be PCR amplified to reach the 1 µg input requirement of ONT library preparations. In contrast, R2C2 is a multistep protocol that, while requiring little hands-on time, is composed of circularization (1 h), linear DNA removal (1–6 h), rolling circle amplification (overnight), and debranching (2 h) followed by size-selection (1 h). However, R2C2 uses only off-the-shelf reagents and requires no special equipment, meaning that performing a pilot experiment to establish whether R2C2 would be superior to 1D reads and a good replacement for any particular short-read assay should be possible for the vast majority of molecular biology laboratories.

Pilot experiments might be required for new library types because converting short-read libraries with R2C2 and sequencing them on an ONT sequencer may change what molecules in a pool will be sequenced. This is a consequence of R2C2 requiring several processing steps and ONT sequencers featuring a unique

underlying technology that is totally distinct from Illumina or any other short-read sequencing technology. For example, in some experiments, R2C2/ONT sampled longer molecules than Illumina sequencers. Further, in the ChIP-seq experiment alone, those longer reads were also more GC-rich. Additionally, applications in which very high read and/or consensus accuracy is required, for example, somatic variant calling, will pose a challenge for R2C2. In essence, before R2C2 is used for a short-read experiment, the requirements for this experiment should be carefully considered.

In addition to Illumina libraries, the R2C2 method can also be easily adapted to libraries generated for one of several other sequencing instruments now entering the market, simply by modifying the splint used to circularize the library. As part of our C3POa tool, we now provide a script that designs splints and the oligos needed to make them for any amplified sequencing library based on the primers used to amplify it.

Beyond simply competing with benchtop sequencers like the Illumina MiSeq, R2C2 can be used for a new group of assays around “medium-length” 600- to 2000-nt reads. Libraries with insert lengths of this size can be size-selected from standard Illumina library preparations, and R2C2 is easily adapted to libraries with different insert lengths by modifying the size-selection of its rolling circle amplification product to include only molecules bigger than three to four times the original library size. We provided one example of the resulting “medium-length” R2C2 reads by analyzing size-selected Tn5 libraries. We showed that these reads can, for example, provide an advantage for the sequencing of small genomes. Among many other potential applications, “medium-length” reads could be applied to standard fragmentation-

based RNA-seq libraries to provide more contiguous splicing information for very long transcripts (>15 kb) where full-length cDNA-based approaches fail.

One of the unique strengths of ONT-based sequencing methods is that, beyond the standard approach of analyzing sequencing runs once they are completed, many library metrics can be derived in real-time. This is starting to get exploited in clinical and metagenomics assays with tools like SURPIrt (Gu et al. 2021) or with more powerful tools like MinoTour (Munro et al. 2021). The PLNK tool we developed here is therefore a powerful tool to monitor sequencing runs and can be used for the rapid evaluation of library metrics. This makes it possible to stop a run once a predetermined target coverage is reached or once it is clear whether a library construction and pooling was successful. For example, using PLNK, we showed that key metrics of the RNA-seq, ChIP-seq, and enriched Tn5 libraries can be evaluated in under 1 h of sequencing, making it possible to flush, store, and reuse the flow cells used for these experiments.

In summary, we have shown that, using R2C2, the ONT MinION can—with some limitations—be used as an accurate short-read sequencer with several advantages over dedicated short-read sequencers. Because the ONT MinION comes with minimal instrument cost, R2C2 allows standard short-read genomic assays to be performed in any laboratory immediately after a library is produced. The use-cases for this, just as the many use-cases for Illumina benchtop sequencers, will vary from laboratory to laboratory. For laboratories performing small-scale experiments—like RNA-seq of a few samples—the R2C2/ONT MinION combination should be entirely sufficient. For laboratories performing largescale experiments—like ChIP-seq of dozens of samples—the R2C2/ONT MinION combination should be

useful to rapidly evaluate library pool compositions and metrics before committing to the cost and turnaround time that deeply sequencing a library pool at a core facility on an Illumina HiSeq or NovaSeq 6000 requires.

In either case, the presence of a capable short-read sequencer in most molecular biology or clinical laboratories could be truly disruptive by eliminating long turnaround times and therefore accelerating experiments.

Methods

Library Preparation

RNA-seq

Four RNA-seq libraries were prepared with the NEBNext Ultra II directional RNA library prep kit for Illumina (NEB E7760) following the manufacturer's protocol. For each library, 100 ng of poly(A) selected RNA from the human lung carcinoma cell line A549 (Takara 636141) was used as input. The RNA fragmentation step was performed for 5 min at 94°C. PCR enrichment of adaptor ligated DNA was performed for nine cycles using the NEBNext multiplex oligos for Illumina (NEB E7600S) kit to add Illumina dual index sequences. Three libraries were pooled at a 4 ng, 2 ng, and 1 ng before sequencing on an Illumina MiSeq instrument for paired-end 2 Å~ 300-bp sequencing. The same three RNA-seq libraries were pooled again at the same ratio for further R2C2 library preparation. For the 1D and R2C2 runs, the fourth RNA-seq library was prepared and added right before ONT library preparation.

ChIP-seq

ChIP was performed following the detailed protocol of Ricci et al. (2020) with minor modification. In brief, approximately 30 developing seeds at the cotyledon stage were used for chromatin extraction. Immediately after harvesting, the tissue was crosslinked as described in the referenced protocol and immediately flash-frozen in liquid nitrogen. To make antibody-coated beads, 25 µL Dynabeads Protein A (Thermo Fisher Scientific 10002D) were washed with ChIP dilution buffer and then incubated with 2 µg antibodies (anti-H3K4me3, Millipore-Sigma 07-473) for at least 3 hours at 4°C. After the nuclei extraction, the lysed nuclei suspension was sonicated to 200–500 bp on a Diagenode Bioruptor on the high setting for 30 min. Tubes were centrifuged at 12,000g for 5 min at 4°C, and the supernatant was transferred to new tubes. At this point, 10 µL of ChIP input aliquots was collected. Sonicated chromatin was diluted 10-fold in the ChIP dilution buffer to bring the SDS buffer concentration down to 0.1%. The diluted chromatin was incubated with antibody-coated beads overnight at 4°C, washed, and reverse-crosslinked. The library was prepared in accordance with the referenced protocol.

Target-enriched Tn5

The Tn5 library was prepared using genomic DNA from cell lines NCI-H1650 (ATCC CRL-5883D) and NCI-H1975 (ATCC CRL- 5908DQ). A total of 100 ng genomic DNA of each sample was treated with Tn5 enzyme loaded with Tn5ME-A/R and Tn5ME-B/R. The Tn5 reaction was performed using 1 µL of the gDNA, 1 µL of the loaded Tn5-AR, 1 µL of the loaded Tn5-BR, 13 µL of H2O, and 4 µL of 5Å~ TAPS-PEG buffer and incubated for 8 min at 55°C. The Tn5 reaction was inactivated by cooling down to 4°C and the addition of 5 µL of 0.2% sodium dodecyl sulfate and then

incubated for 10 min. Five microliters of the resulting product was nick-translated for 5 min at 72°C and further amplified using KAPA Hifi Polymerase (KAPA) using Nextera_Primer_B_Universal and Nextera_Primer_A_Universal (Smart-seq2) with an incubation for 30 sec at 98°C, followed by 16 cycles of (20 sec at 98°C, 15 sec at 65°C, 30 sec at 72°C) with a final extension of 5 min at 72°C.

The resulting Tn5 library was then enriched with Twist fast hybridization reagents and customized oligo panels that were designed based on the Stanford STAMP panel. The hybridization reaction of the panel and the Tn5 libraries was performed using 294 ng of NCI-H1975 Tn5 library, 360 ng of NCI-H1650 Tn5 library, 8 µL of blocking oligo pool (100 µM), 8 µL of universal blockers, 5 µL of blocker solution, and 4 µL of the custom panel. The mix was dehydrated using a SpeedVac and was resuspended in 20 µL fast hybridization mix at 65°C. After the addition of 30 µL of hybridization enhancer, the mixture was incubated for 5 min at 95°C and for 4 h at 60°C. After hybridization, the reaction mix was incubated with prewashed streptavidin binding beads and washed using the fast wash buffer one and fast wash buffer two for six times. The streptavidin beads and the DNA mixture were used directly for reamplification with universal primers and Equinox library amp mix. The mixture was incubated for 45 sec at 98°C, followed by 16 cycles of (15 sec at 98°C, 30 sec at 65°C, 30 sec at 72°C) with a final extension of 1 min at 72°C. The final enriched Tn5 library DNA product was cleaned up using SPRI beads at 1.8:1 (beads:sample) ratio.

R2C2 conversion

Pooled Illumina libraries were first circularized by Gibson assembly with a DNA splint containing end sequences complementary to ends of Illumina libraries. Illumina

libraries and DNA splint were mixed at a 1:1 ng ratio using NEBuilder HiFi DNA assembly master mix (NEB E2621). Any non-circularized DNA was digested overnight using ExoI, ExoIII, and Lambda exonuclease (all NEB). The reaction was then cleaned up using SPRI beads at a 0.85:1 (bead:sample) ratio. The circularized library was then used for an overnight RCA reaction using Phi29 (NEB) with random hexamer primers. The RCA product was debranched with T7 endonuclease (NEB) for 2 h at 37°C and then cleaned using a Zymo DNA Clean and Concentrator column-5 (Zymo D4013). The cleaned RCA product was digested using NEBNext dsDNA fragmentase (NEB M0348) following the manufacturer's protocol with a 10-min incubation. For the regular Tn5 library, digested RCA product was cleaned using SPRI beads. For all other libraries, the digested RCA product was size-selected using a 1% low-melt agarose gel: DNA between 2 and 10 kb was excised from the gel, which was then digested using NEB beta-agarase. DNA was then cleaned using SPRI beads.

ONT sequencing

ONT libraries were prepared from R2C2 DNA or directly from Illumina libraries using the ONT ligation sequencing kit (ONT SQK-LSK110) following the manufacturer's protocol and then sequenced on an ONT MinION flow cell (R9.4.1). When preparing ONT libraries from Illumina libraries, SPRI bead purifications throughout the protocol were adjusted to accommodate for their short length. Additional library was loaded on the same flow cell after nuclease flush.

Analysis

Real-time analysis with PLNK

RNA-seq, ChIP-seq, and Enriched Tn5 MinION runs were simulated by reading the mtime metadata entry of FAST5 files in the output folder of the completed runs and then calculating the time intervals at which files were created by the MinKNOW software. Files created during the first 48 h or until the first library reload were then copied into a new folder at those intervals. PLNK (https://github.com/kschimke/PLNK) was started after the simulation and was given key information about the run (splint and Illumina indexes in the format of a sample sheet, target regions in BED format, genome sequence in FASTA format) and a config file containing paths to tools used by PLNK.

General analysis

SAMtools (v1.11-18-gc17e914) (Li et al. 2009) was used extensively during analysis for SAM file processing. Python (Oliphant 2007), Matplotlib (Hunter 2007), Numpy (Harris et al. 2020), and SciPy (Virtanen et al. 2020) were all used to analyze and visualize the data.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA775962. All codes used for analysis are available at GitHub (https://github.com/kschimke/PLNK; https://github.com/rvolden/C3POa).

Generating cell type specific transcriptomes for human lung tissue

Introduction

In the same way that we could transform Tn5, ChIP-seq, and RNA-seq Illumina libraries into R2C2 product for ONT sequencing, previous experiments used R2C2 in combination with 10X Genomics single cell cDNA pools (Volden and Vollmers 2022). 10X pools are composed of full length cDNA with added cell barcodes which is usually fragmented for short read sequencing on an Illumina sequencer. While short read sequencing produces highly accurate reads which in turn increases cellular identification per molecule, fragmentation limits our ability to identify variability between transcripts because of a bias towards the 3' end preventing most of the molecule from being sequenced. Transforming 10X pools into R2C2 product before ONT sequencing, not only increases the accuracy and identity of the reads, but also captures full length transcripts. In this study, we have applied the R2C2 protocol to 10X single cell cDNA pools from the Tabula Sapiens Consortium.

The Tabula Sapiens Consortium set out to produce a multi-organ, single cell atlas of humans (Tabula Sapiens Consortium 2022). Their data consists of nearly 500,000 cells from 24 organs of 15 normal human subjects. Their donors have a mean age of 51, are balanced by gender, and represent a range of ethnicities and medical backgrounds. They put in the effort to ensure that tissue samples were collected and processed from each donor within a day to prevent degradation and enforce consistency across donors. For each donor and tissue, they performed SmartSeq2 and/or 10X Genomics single cell processing followed by Illumina

sequencing. Their main goal was to create a comprehensive single cell dataset that would be broadly available for the scientific community.

Results

Since the Tabula Sapiens database contains only short read data, we focused on sequencing their 10X pools using our R2C2 protocol in combination with ONT sequencing to produce a robust long read dataset (Fig. 4A). We received 10X Genomics single cell cDNA pools from lung tissue of two of the Tabula Sapiens Consortium donors (TSP2 and TSP14). The two samples were handled separately during R2C2 preparation, each receiving a different DNA splint, before being pooled together and processed with ONT's library preparation kit. The library was sequenced over 12 runs on three ONT PromethION flow cells with DNAse flushes between runs. After basecalling with guppy5, the dataset contained 99,086,987 R2C2 long reads which, after processing and demultiplexing with C3POa, resulted in 49,649,110 consensus reads for TSP2 and 30,906,060 consensus reads for TSP14. We compared the R2C2 raw read length with the resulting insert lengths to ensure that our long reads had sufficient coverage of the original input library (Fig. 4B).

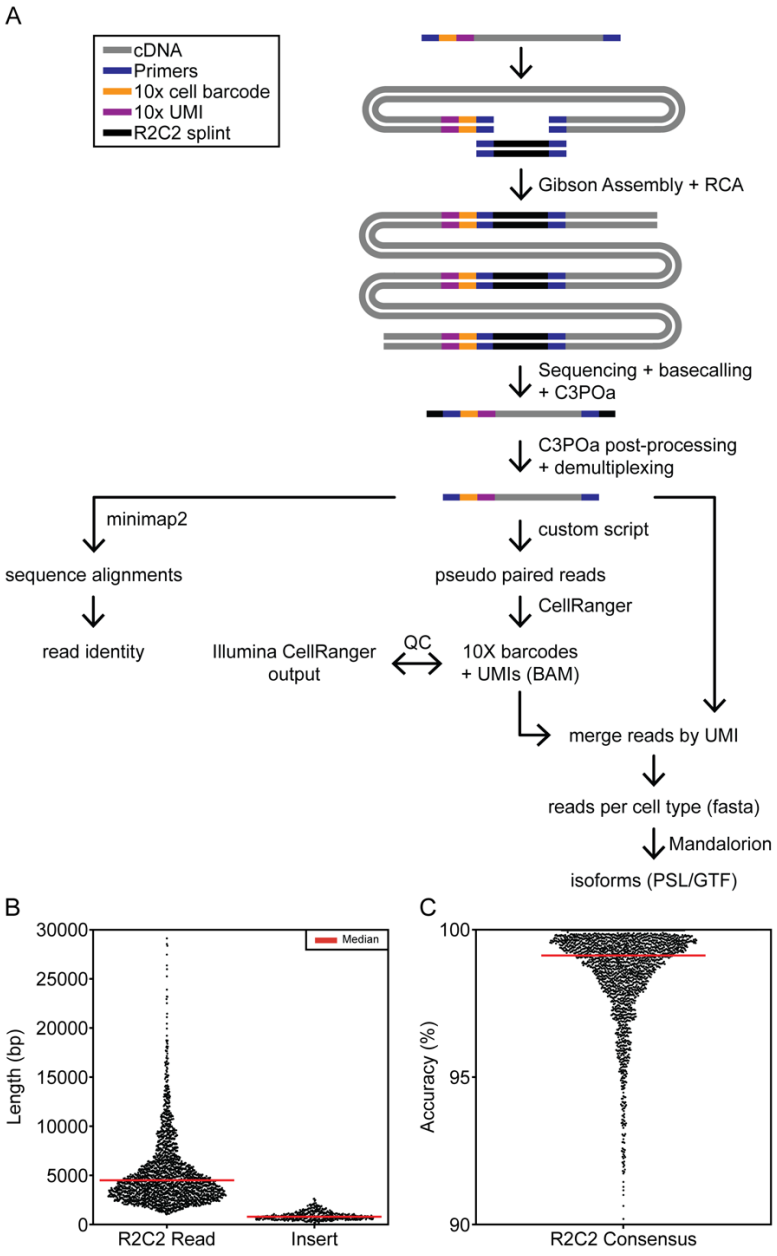

Figure 4. R2C2 10X single cell preparation and analysis overview. (A) Molecules from a 10X single cell library are circularized with Gibson assembly and amplified during RCA. The long R2C2 product is then sequenced on an ONT instrument, basecalled with guppy5 and consensus called with C3POa. The R2C2 library is demultiplexed with C3POa and the resulting 10X reads are transformed into pseudo paired reads and analyzed with Cell Ranger to identify barcodes and UMIs. The reads are merged by 10X UMI and then sorted into cell types for isoform analysis with Mandalorion. (B) R2C2 raw read

length and insert length comparison. The median raw read length was 4515 bp, median insert length was 805 bp. (C) R2C2 consensus read accuracy was determined by aligning reads to the human reference genome. Median read accuracy is 99.15%.

Along with the 10X library pools, the Tabula Sapiens Consortium also provided their Illumina sequencing data with cell barcodes analyzed by Cell Ranger and cell type annotations for highly ranked cell barcodes from TSP2 and TSP14. To directly compare our long read data with their Illumina data, we had to demultiplex the 10X cell barcodes with Cell Ranger, a software package designed for Illumina read pairs. Since Cell Ranger expects Illumina reads, we needed to assess the accuracy of our R2C2 ONT data to determine if the quality of the data is sufficient enough to match Illumina read quality. We aligned the R2C2 consensus reads to human genome reference using minimap2 (Li 2018) and found that our median read accuracy was 99.15% which is comparable to Illumina (Fig. 4C). To make our ONT sequencing output compatible with Cell Ranger, we developed a python script which creates mock Illumina read1 and read2 FASTQ files from ONT reads. We then ran Cell Ranger with the mock FASTQ files to identify the 10X cell barcode and UMI barcode associated with each ONT read. Reads containing matching UMI barcodes originated from the same RNA molecule in the original sample and were merged using a python script resulting in <10% of reads merged for both donors.

The Tabula Sapiens Consortium determined which barcodes to annotate with cell types based on the barcode's frequency within their short read dataset. We found that the number of reads assigned to each barcode by Cell Ranger correlates between the Illumina dataset and our ONT dataset (Fig. 5CD) and that a majority of

the barcodes annotated by the Tabula Sapiens Consortium fall within the highest frequency barcodes from our ONT dataset (Fig. 5AB).

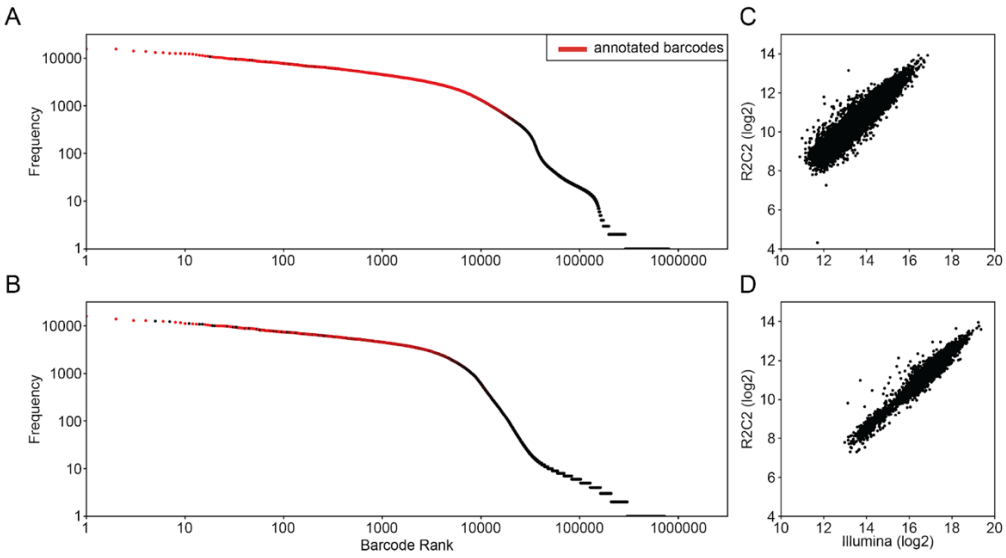

Figure 5. Comparison of barcode assignment in R2C2 10X data and Illumina 10X data. Given the annotation data for the cells identified by the Tabula Sapiens Consortium as well as the raw Illumina reads analyzed with Cell Ranger we can quality check our R2C2 data for TSP2 (AC) and TSP14 (BD). (AB) Barcode rank is determined by the number of reads assigned to a given cell barcode by Cell Ranger. Cell barcodes denoted in red were annotated by Tabula Sapiens because they have a high barcode rank in the Illumina data. (CD) A comparison of the number of reads from the R2C2 and Illumina datasets that were assigned to the cell barcodes annotated by Tabula Sapiens.

After validating the cell composition of the R2C2 data, we used Mandalorion for isoform analysis (Volden et al. 2023). We referenced the annotation data provided by the Tabula Sapiens Consortium to match cell barcodes with cell types and then sort the R2C2 reads into cell type FASTA files based on the cell barcodes assigned by Cell Ranger. Previously, we treated TSP2 and TSP14 separately but here we combined the data from the two donors based on cell type to increase the number of reads per cell type. We ran Mandalorion on each cell type individually to

find cell type specific isoforms (Fig. 6), as well as on the dataset as a whole to compare isoform usage between cell types (Fig. 7). We followed up each Mandalorion run with SQANTI3 which classified the isoforms by comparing them to a reference transcriptome (Tardaguila et al. 2018). We see a significantly higher percentage of incomplete splice match isoforms within the R2C2 10X data when compared with Mandalorion output from a normal R2C2 cDNA library. We find about 10% incomplete splice match in bulk RNA analysis but see >25% incomplete splice match in the most abundant cell types for this dataset. This could be due to the nature of Illumina libraries which favor the 3'-end and would naturally capture fewer full length transcripts.

SQUANTI3 also identified the genes associated with each isoform which we combined with the quantification data from Mandalorion to be able to determine differential isoform usage within genes. We performed a chi-square test on the Mandalorion output from the entire dataset and found the five most significant genes based on differential isoform expression are as follows: S100A16, MYL6, SERPINF1, LIMS1, and RPS24. The Tabula Sapiens Consortium identified MYL6 in their 2022 paper because of an exon skipping event that occurs in different groups of cell types. They determined that the sixth exon is skipped in immune and epithelial cell types, while stromal and endothelial cell types include the exon. We were able to corroborate their findings for MYL6 with the R2C2 dataset further validating the accuracy of the long read data (Fig. 7A).

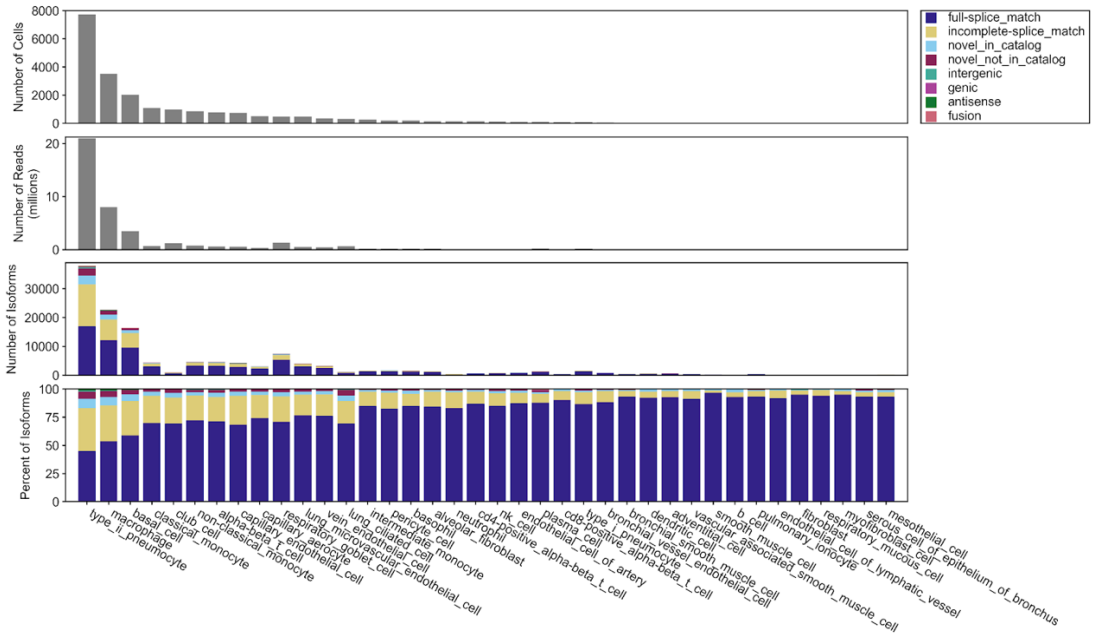

Figure 6. Isoform classification of cell types in R2C2 10X data. Combining R2C2 ONT read data from both TSP2 and TSP14, we show (1) the number of cells, (2) the number of reads, (3) the number of isoforms for each cell type within specific categories, and (4) the percentage of isoforms within each category for each cell type.

The Tabula Sapiens Consortium also analyzed alternative splice sites in CD47 isoforms. They concluded that immune, endothelial and stromal cell types favor 3' splice sites while epithelial cell types favor exons closer to the 5' splice site. Again, we came to the same conclusions using the CD47 isoforms produced by Mandalorion (Fig. 7B). While corroborating the Tabula Sapiens Consortium's conclusions validates our sequencing protocol and analysis pipeline, we know that there's more that can be discovered from our data. Further inspection of the transcriptomes produced by Mandalorion from our R2C2 10X dataset could improve the community's understanding of key genes.

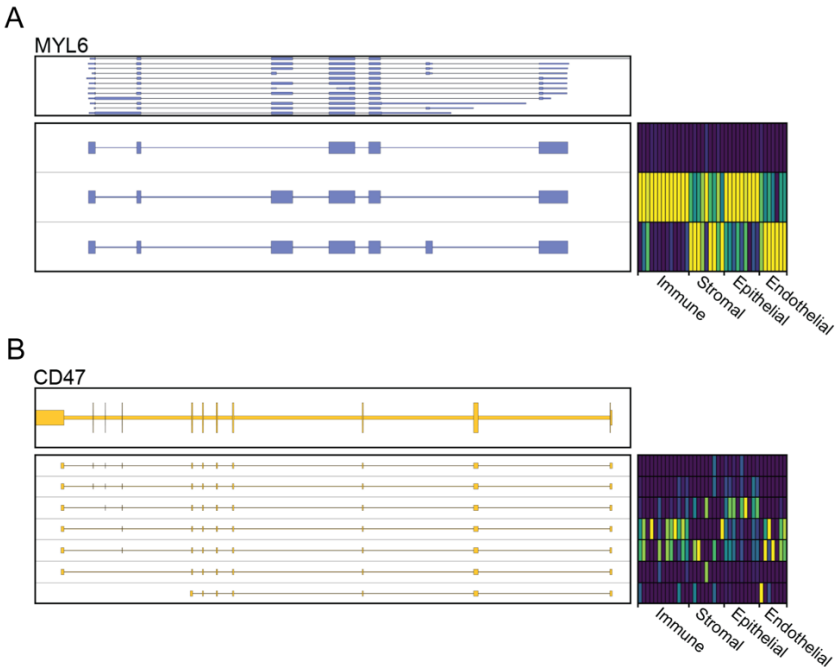

Figure 7. Differential isoform expression for MYL6 and CD47. For both genes (1) the reference isoforms as defined by GENCODE, (2) the isoforms produced by Mandalorion from the R2C2 10X single cell data and (3) a heatmap of the normalized expression value for cell types within tissue compartments where yellow is highly expressed and purple is no expression. (1-2) Isoforms in blue map to the positive strand while those in yellow map to the negative strand, both oriented with the 5'-end on the left. (A) An exon skipping event in MYL6: the exon is skipped in immune and epithelial cell types and kept in stromal and endothelial cell types. (B) Alternative splicing near the 5'-end, epithelial cell types tend to splice in exons near the 5'-end while immune, stromal and endothelial cell types favor 3' splice sites.

Conclusion

We first applied R2C2 to 10X Genomics single cell libraries in 2022 (Volden and Vollmers 2022) and successfully classified and quantified isoforms with that data. This study demonstrates a marked improvement of the R2C2 and 10X single cell protocol. Where before the comparatively low accuracy of R2C2 reads required us to

develop software for demultiplexing the 10X libraries, we can now produce R2C2 reads at an accuracy that is comparable to Illumina, allowing us to use commercial software for demultiplexing. From our demultiplexed data, we determined that we sequence the same distribution of molecules within the 10X libraries as the Illumina dataset provided by the Tabula Sapiens Consortium. We found that we had similar ratios for the number of reads assigned to each 10X barcode when comparing the Illumina dataset with our R2C2 dataset. Finally, we determined that the R2C2 dataset is just as powerful as the Illumina dataset when analyzing isoforms by comparing the isoform expression levels of specific genes from the Tabula Sapiens study.

While the R2C2 10X dataset generated in this study is substantial and informative, we can improve the isoform analysis by also studying bulk RNA samples. Initial sequencing runs with bulk lung RNA using R2C2 show improved raw read length and increased median insert sizes. An increased read length means there is a higher likelihood of full splice matches when analyzing isoforms which could fill the gap left by the limited length of the R2C2 10X reads. Matching our R2C2 10X reads with a bulk sequencing run will enable us to polish the cell type specific isoforms producing more complete transcriptomes.

We would like to thank the Tabula Sapiens Consortium for providing us with their 10X library pools, Illumina sequencing data and single cell analysis.

Sequencing full-length plasmids with ONT long-reads

This section is adapted from **Sequencing full-length plasmids with ONT long-reads using R2C2** (Schimke and Vollmers 2024)

Sequencing full-length plasmids with ONT long-reads using R2C2

Kayla D. Schimke[1] and Christopher Vollmers[1,*]
1) Department of Biomolecular Engineering,
*) Corresponding author. Email: vollmers@ucsc.edu

Abstract

Plasmids are robust tools in molecular biology and are often utilized for various experiments within a traditional lab. Overtime and with multiple transformations, the base structure of a plasmid can change and potentially hinder the function of the plasmid. Typically, underperforming or unknown plasmids undergo sequencing analysis to determine whether or not they are functional, oftentimes only verifying the insert and ignoring the base structure. We adapted the R2C2 sequencing protocol for plasmids and developed an analysis pipeline, Chopper, that produces full length assemblies of the input plasmids. We tested our protocol and software with pure plasmid samples from addgene, sequencing the plasmids on their own and within a pool with ONT PromethION sequencing. We produced highly accurate assemblies for each plasmid from both their individual sequencing run and from the pool.

Introduction

Plasmids are a fundamental tool in molecular biology as they are vectors for recombinant DNA. Their role warrants validating their composition, usually through sequencing technologies, to ensure they retain their functionality over many generations. Plasmid sequencing has been dominated by Sanger sequencing and short read sequencing using Illumina technologies. Multiple institutions, including Applied Biological Materials Inc, seqWell, Massachusetts General Hospital Center for Computational and Integrative Biology, and CD Genomics, offer plasmid sequencing on Illumina platforms. While short read sequencing is highly accurate, the turnaround time is less than ideal for quick validations (>2 weeks) and the reliance on fragmenting before sequencing could prevent complete assembly. Some of the previously mentioned institutions charge a fee for assembling and annotating a plasmid sample and still cannot guarantee complete assemblies. A few sequencing facilities have turned to long read sequencing to ensure they capture full length plasmids; CD Genomics offers sequencing with Pacific Biosciences (PacBio) SMRT technology specifically for large plasmids, while eurofins and Plasmidsaurus sequence plasmids exclusively with Oxford Nanopore Technologies (ONT). These facilities offer a quicker turnaround time, but their analysis is still limited by the accuracy of long read sequencing.

ONT sequencing makes it possible to sequence in the lab, using a MinION or P2 Solo, while capturing full length plasmids with long reads. While updates to their protocol and software continue to improve the accuracy of their reads, ONT sequencing still falls short of Illumina when comparing median read accuracy. In an effort to make the accuracy of long reads comparable to that of short reads,

researchers in the Vollmers lab at UCSC developed the Rolling Circle Amplification to Concatemeric Consensus (R2C2) protocol to transform input DNA before ONT sequencing. This protocol involves circularizing linear DNA with an identifying splint (UMI) using Gibson assembly, before performing rolling circle amplification (RCA) to produce molecules containing multiple copies of a single template (Volden et al. 2018). We aim to prove that using a modified version of the R2C2 protocol and sequencing in lab on an ONT P2 Solo would be more time and cost effective than previous methods or sequencing centers while still producing highly accurate full length plasmids.

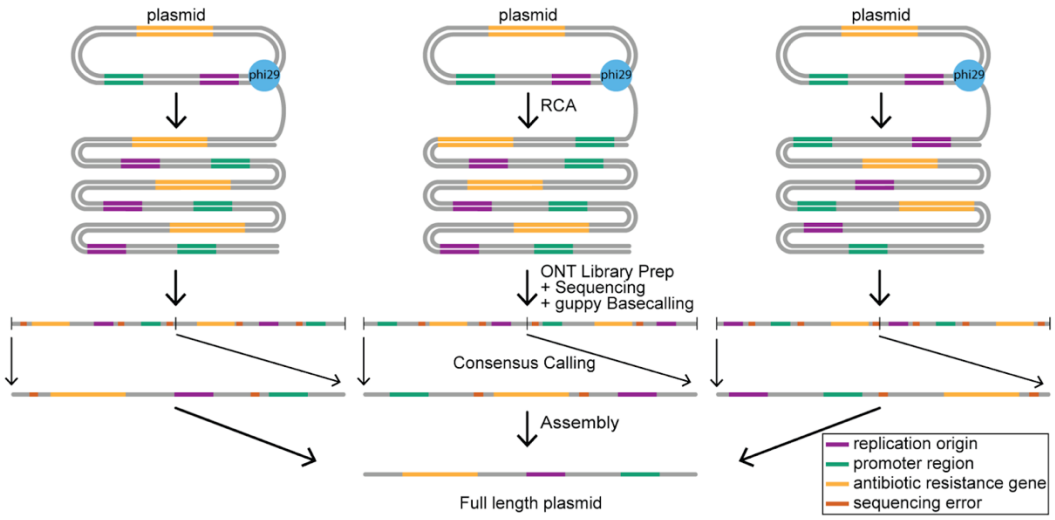

Figure 8. Modified R2C2 for plasmid sequencing. The first step in the R2C2 protocol involves using Gibson assembly to circularize DNA/cDNA molecules with a UMI. Since plasmids are by nature circular, we eliminated the circularization step and immediately began rolling circle amplification (RCA) to produce linear DNA containing multiple copies of the original plasmids. Three copies of the same plasmid are shown before and after RCA to demonstrate that the same plasmid can produce different linear molecules depending on where phi29 binds during replication. The linear DNA was prepped and sequenced on an ONT PromethION and then consensus called before assembling the reads into full length plasmids.

Plasmids make the perfect input for R2C2 as they are already circles. By omitting the Gibson assembly step from R2C2 and instead just amplifying the plasmids with Phi-29, we produced long DNA strands containing multiple copies of the plasmids without the defining UMIs (Fig. 8). Since the modified R2C2 molecules lack UMIs, we cannot use C3POa, the accompanying consensus calling software for R2C2, in its current form. C3POa expects molecules with UMIs because it aligns the UMI to the molecule to determine the location of repeats. As an alternative method for consensus calling, TideHunter (https://github.com/yangao07/TideHunter) was developed to detect tandem repeats in long read sequencing using a seed and chain algorithm (Gao et al. 2019). TideHunter was designed for and tested with ONT and PacBio data with error rates of up to 20% making it perfect for R2C2 reads sequenced on a P2 Solo.

Results

To test the modified R2C2 protocol and analysis pipeline, we ordered a set of four plasmids with known references from addgene: pCSCMV:tdTomato, lentiCRISPR v2, pSpCas9(BB)-2A-Puro (PX459) V2.0, pcDNA3.1-GFP(1-10). Each plasmid was prepped individually and as part of a pool containing equal amounts of each plasmid. After RCA, we performed a gel size selection with the R2C2 product, keeping only DNA that was over 10 kb. The R2C2 product for each plasmid and the pool was then library prepped and sequenced on an ONT P2 Solo. For these experiments we purposely sequenced on PromethION flow cells that had been used for previous experiments and limited the run to about an hour. Since we expect to see full length plasmids within our reads, we need fewer reads to get sufficient coverage for each

plasmid. Because flow cell pore retention between DNAse flushes is relatively high, we get enough reads within an hour run that a used flow cell should not affect the accuracy of our final assembly.

Consensus calling modified R2C2 data

After sequencing each individual plasmid and the pool, we basecalled the reads with guppy5 (dna_r10.4.1_e8.2_400bps_sup.cfg) and obtained 58,209 reads, 22,626 reads, 232,970 reads, 136,012 reads and 193,117 reads for pCSCMV:tdTomato, pcDNA3.1-GFP(1-10), lentiCRISPR v2, pSpCas9(BB)-2A-Puro (PX459) V2.0, and the pool respectively. The discrepancy in read numbers is directly related to the state of the flow cell before the experiment; the lower the number of viable pores on the flow cell resulted in fewer reads. Each set of reads was consensus called with TideHunter. Since we expect the consensus reads to contain full length plasmids, a majority of the consensus reads when binned by length should accumulate into bins corresponding with the actual lengths of the plasmids creating visible peaks (Fig. 2A-E). This was true for all the addgene plasmids except for lentiCRISPR v2 when processed by TideHunter (Fig. 2C).

TideHunter by default expects a minimum of two repeats of an insert within a raw read. The lentiCRISPR v2 plasmid is about 15 kb, meaning that our R2C2 reads would need to be about 30 kb in length to see two full repeats of this plasmid. As shown in the top panel of figure 2C, our raw read length peters off around 8 kb with a small peak at 15 kb, around the reference length of the lentiCRISPR v2 plasmid. A longer sequencing run could have yielded enough reads around 30 kb for

TideHunter to detect the full length plasmid but given the peak in the raw reads we know we have enough data to make a complete consensus just not with TideHunter.

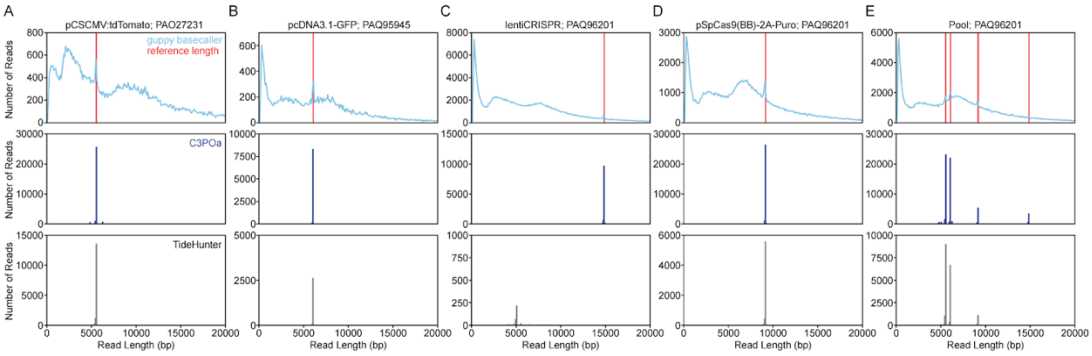

Figure 9. Comparison of C3POa and TideHunter consensus reads. (A-E) The read length distributions of the raw sequencing data (top), C3POa consensus (middle), and TideHunter consensus (bottom). The length of the addgene reference sequences for each plasmid are shown as red bars highlighting that consensus reads pile up around their plasmid lengths.

C3POa, Concatemeric Consensus Caller with Partial Order alignments, is the companion software for the R2C2 protocol (Volden et al. 2018). As stated before, this software was designed to create consensus reads from R2C2 molecules which have a specific structure, where known sequences mark repeats of input molecules. The absence of a known sequence delineating repeats in our plasmid R2C2 data initially deterred us from using C3POa for consensus calling and led us to TideHunter. The inability for TideHunter to create consensus reads for our longest plasmid encouraged us to adapt C3POa for the modified R2C2 data.

We introduced a new option for C3POa, --nosplint or -ns, which assumes that the raw data does not contain a UMI or other known sequence. In this case, instead of aligning a user defined known sequence to each raw read to determine the location of repeats we would have to define a sequence that is unique to each read.

Initially we attempted to align the first 300 bp of the read to itself with varying degrees of success. Namely, the pCSCMV:tdTomato plasmid contains a tandem repeat of the dTomato gene which is the main functional unit of the plasmid. Because of this, we found that aligning the very beginning of a read to itself resulted in two peaks within the consensus reads, one matching the length of dTomato and another matching that of the pCSCMV:tdTomato reference. We determined that using a 200 bp region further into the read eliminated the second peak within the pCSCMV:tdTomato consensus data and increased the number of consensus reads across all datasets.

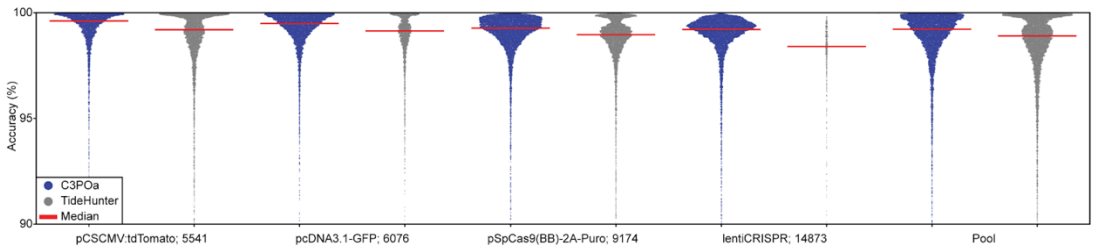

Figure 10. Evaluation of C3POa and TideHunter consensus accuracy. Each sequencing run for the individual plasmids is labeled with the plasmid name and reference length to highlight the correlation between median read accuracy, shown as red bars, and plasmid length.

Once we had consensus sequences from all datasets using TideHunter and the updated C3POa, we evaluated their accuracy by aligning the output of both tools to the addgene reference using minimap2 (Li 2018) (Fig. 10). Accuracy was calculated by dividing the number of matches by the sum of matches, mismatches and indels within the read. The median consensus accuracy for the pool using C3POa was 99.21% (99.61% pCSCMV:tdTomato, 99.49% pcDNA3.1-GFP(1-10), 99.20% lentiCRISPR v2, 99.26% pSpCas9(BB)-2A-Puro (PX459) V2.0) versus using

TideHunter was 98.90% (99.19% pCSCMV:tdTomato, 99.14% pcDNA3.1-GFP(1-10), 98.40% lentiCRISPR v2, 98.96% pSpCas9(BB)-2A-Puro (PX459).

In general, as the length of the plasmids increased the median accuracy decreased, but C3POa still outperformed TideHunter. The considerably high accuracy of the TideHunter output for lentiCRISPR v2 proves that while the consensus reads don't contain the full plasmid, it was still able to recover partial fragments. We also noticed that the distributions of the accuracies for TideHunter appear bimodal which could be a direct result of its strict repeat definitions. TideHunter only considers full insert repeats when creating a consensus. C3POa, on the other hand, includes partial inserts when creating consensus reads which contributes to its more unified distributions (Fig. 10).

Assembling plasmids with Chopper

After consensus calling, we wanted to assemble our reads to generate a single output for each plasmid to better assess the content of the plasmid. We performed a comparative analysis of assembly tools with applications specifically for plasmids. We only used the sequencing data from the pool for this analysis because, of our five datasets, it is the most robust. We compared the output from metaFlye (Kolmogorov et al. 2020) and Miniasm (Li 2016). We excluded Unicycler (Wick et al. 2017) from our analysis because it utilizes Miniasm within its pipeline. The two metrics we considered were the number of contigs the tool produced as well as the accuracy of the assembly, assessed by aligning the assembly to the addgene reference using minimap2.

Since all four addgene plasmids are present in the pool, we expect the assemblers to produce four contigs each composed of only one of the plasmids. MetaFlye and Miniasm were given the complete plasmid pool dataset for assembly and were run using the optimal settings for plasmids as defined by their developers. MetaFlye produced four contigs while Miniasm produced 21 contigs. Each contig from metaFlye was at least double the length of its corresponding plasmid. After aligning the assembly to the addgene reference, we found that each metaFlye contig contained an individual plasmid but with large regions of extraneous bases on either side of it. Miniasm produced too many contigs but the length of each was more reasonable for the plasmids sequenced. Aligning the contigs to the addgene reference revealed that Miniasm wasn't able to assemble any of the plasmids in full.

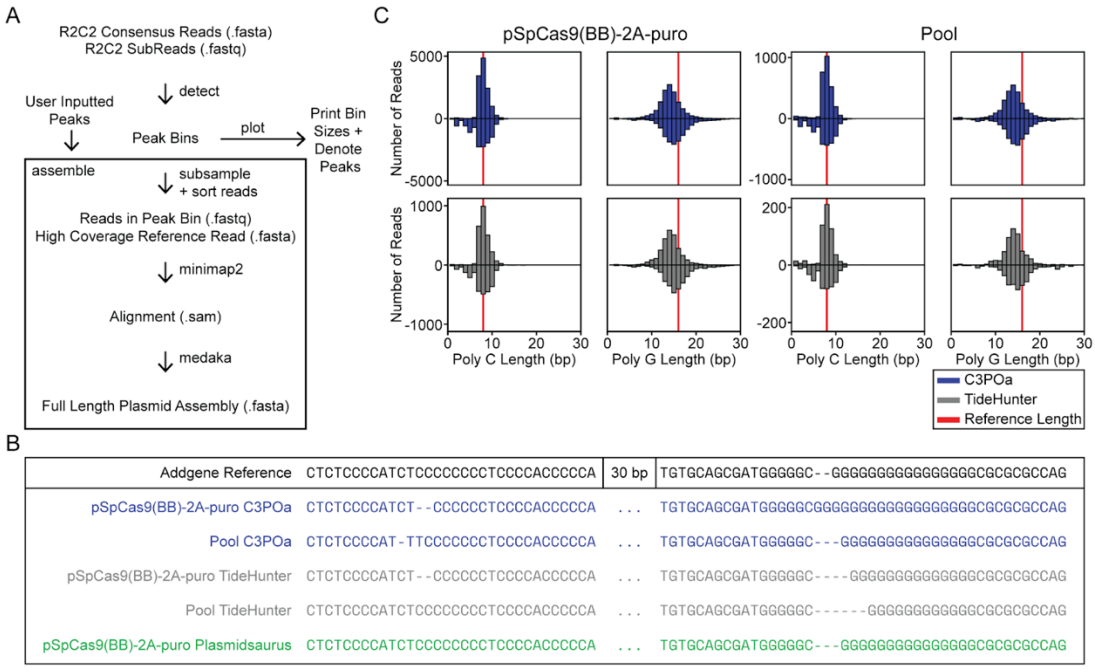

Figure 11. Assembling plasmids with Chopper. (A) Diagram of Chopper functionality; three modes that can be run in succession or individually: detect, plot, and assemble. Detect: bins reads from C3POa (both consensus FASTA and subread FASTQ) or TideHunter (only consensus FASTA) output and

determines peak bins. Assemble: utilizes minimap2 and medaka consensus to assemble the reads within the detected bins or at user defined peaks. Plot: prints a visual representation of the bins with the detected peaks labeled. (B) Multiple alignment of Chopper assemblies for pSpCas9(BB)-2A-puro from the individual and pool sequencing runs using either C3POa or TideHunter for consensus and the individual output from Plasmidsaurus. (C) Distributions of the lengths of the two homopolymer regions from the C3POa and TideHunter consensus reads.

Given that neither of the assemblers produced viable assemblies, we developed our own assembly pipeline called Chopper (Fig. 11A). The software has three modes: detect, plot and assemble. Each of the modes can be run independently of one another using consensus data from either C3POa or TideHunter. The detect mode bins the consensus reads by length and determines peak bins by comparing the size of the bins and finding local maxima. The plot mode runs detect and then prints a range of bins and their sizes, marking any peak bins within the range. The assemble mode processes the peaks found during detect by first creating a reference from a tandem repeat of the highest coverage read within the bin and then aligning the rest of the reads in the bin to that reference using minimap2. Chopper assemble uses that alignment as input for medaka consensus to produce a high accuracy consensus read for the peak which is then trimmed to contain a single full length plasmid.

We applied Chopper to the consensus reads from both C3POa and TideHunter for all five of our sequencing runs. As expected, each run yielded the correct number of contigs for the plasmids present in the sequencing run because Chopper utilizes binning to find the reads that correspond with the lengths of the plasmids. As stated previously, we expect our consensus reads to contain full length

plasmids thus the consensus reads accumulate around the length of the plasmids present in the sequencing run. We evaluated the accuracy of the Chopper assemblies by aligning its output to the plasmid references from addgene and used the same algorithm to calculate identity as we did for the consensus reads. The accuracies for the assemblies from all sequencing runs using either C3POa or TideHunter are displayed in Table 1.

Sequencing Run	Consensus Tool	pCSCMV:tdTomato	pcDNA3.1-GFP(1-10)	lentiCRISPR v2	pSpCas9(BB)-2A-Puro (PX459)
Individual	C3POa	100%	100%	100%	99.96%
Pool	C3POa	100%	100%	100%	99.97%
Individual	TideHunter	100%	100%	NA	99.96%
Pool	TideHunter	100%	100%	NA	99.96%
Individual	Plasmidsaurus	100%	100%	100%	99.989%

Table 1. Chopper assembly accuracies

Most of the assemblies from Chopper were 100% accurate regardless of the consensus tool with two stand outs, lentiCRISPR v2 and pSpCas9(BB)-2A-Puro (PX459). Since TideHunter was unable to make accurate full length consensus reads for lentiCRISPR v2, Chopper wasn't able to detect a peak at that plasmid length and thus couldn't assemble the plasmid from its individual sequencing data nor the pool. We also noticed that no matter the consensus tool, Chopper struggled to assemble pSpCas9(BB)-2A-Puro (PX459) which warranted further analysis of the plasmid structure. Comparing the four Chopper assemblies for this plasmid, we found two homopolymer regions within 30 base pairs of one another that exhibited a high rate of indels (Fig. 11B).

ONT sequencing is known to have issues within homopolymer regions because of the nature of nanopore sequencing. As a strand of DNA moves through

the pore, the instrument records the change in voltage which is interpreted by a basecaller to produce a sequence. The rate at which the DNA strand is pulled through the pore can fluctuate but the voltage is constant so the length of the change in voltage doesn't indicate the length of the DNA sequence. This results in the basecaller misrepresenting the number of bases within a homopolymer region which appears as indels. Looking at the consensus reads from both C3POa and TideHunter for pSpCas9(BB)-2A-Puro (PX459), we counted the number of bases in the two homopolymer regions for each read to determine the distribution of indels (Fig.11C). The shorter of the two regions was less likely to be misrepresented because the length of the region is comparable to the number of bases that span the pore during sequencing. Following this logic, the length of the longer homopolymer is consistently underrepresented by at least 2 base pairs because it is more than double the base span of a pore on the flow cell.

To determine if the issues within the Chopper assemblies were unique or fixable, we sent our plasmid samples to Plasmidsaurus for sequencing and analysis. Plasmidsaurus performs ONT sequencing for multiple types of input, most notably plasmids. They boast overnight sequencing with quick and accurate results. While they did produce a more accurate assembly for pSpCas9(BB)-2A-Puro (PX459) (Table 1), their assembly contained errors in the same region as the Chopper assemblies due to limitations in the sequencing instrument (Fig. 11B), proving that Chopper is robust enough to compete with commercial sequencing and analysis.

Discussion

Plasmids have many applications for molecular biology and often stick around in labs for multiple generations and experiments, constantly undergoing transformations and purifications. It's not uncommon for researchers to find issues with their long standing plasmids from loss of function to unexpected products. The standard for validating plasmids has been Sanger sequencing within the lab but this method lacks the ability to capture full plasmids. Sanger sequencing relies on primers to capture short DNA fragments which limits plasmid validation to the insert region. If the issue with the plasmid lies outside of the insert region, the researcher would have performed Sanger sequencing in vain and could end up discarding the plasmid sample.

Company	**Type of Sequencing**	**Input Requirements**	**Price Per Sample**	**Turnaround Time**
ABM	Illumina paired end	1 - 50 ng	$315*	4-6 weeks
seqWell	Illumina paired end	96 well plate 600 - 1500 ng per well	NA	2 weeks
MGH CCIB	Illumina MiSeq	1400 - 2275 ng	$56.16 - $70.20	5 - 8 business days
CD Genomics	Illumina or PacBio (large plasmids only)	1400 - 2275 ng	$270**	15 business days
eurofins	ONT	300 - 2000 ng **	$15 - $60**	1 business day
Plasmidsaurus	ONT	300 - 2000 ng **	$15 - $60**	1 business day

Table 2. Commercial Plasmid Sequencing Statistics. *additional cost for assembly and annotation **dependent on plasmid size

Instead of verifying only the insert, researchers have the option to turn to commercial whole plasmid sequencing. As shown in Table 2, there are a plethora of institutions that offer whole plasmid sequencing, but the timeframe and price varies dramatically. Companies that perform Illumina sequencing tend to be more expensive with longer turnaround times and require more input material. For other commercial entities these demands may seem negligible but for the average researcher in an academic lab, they can be a strain on their time and resources. Illumina sequencing also does not guarantee a full length plasmid assembly, especially for large plasmids, because the input material is fragmented before sequencing. Because of the potential difficulty with assembly, some companies charge an extra fee for assembly and analysis of the plasmid. Opting to forgo this fee would require the researcher to devote more time to analyzing their plasmid samples after waiting multiple weeks for their sequencing data.

A few companies have turned to long read sequencing, either PacBio or ONT, with the intention of reducing the cost of sequencing and improving their ability to assemble full length plasmids. While long read sequencing lacks the accuracy of Illumina reads, the read length ensures contiguity. The ability of long reads to capture full length plasmids reduces the number of reads needed to get sufficient coverage for assembly, thus companies can decrease their input requirements and reduce their turnaround time with shorter sequencing runs.

Plasmidsaurus, a company that uses ONT sequencing, also boasts the shortest turnaround time at the cheapest cost. They offer convenient drop boxes with weekly sample pickups located on site at several research institutions across the US. Once they receive a sample, they adhere to an overnight sequencing protocol which

provides results the next day. While this may seem quick, they do maintain that samples can take up to a week to arrive at their processing facility. We sent Plasmidsaurus the plasmids we sequenced in this study as a way of verifying our own sequencing results and validating their turnaround time. We received data three days after submitting our samples via their drop box which is a comparable timeframe to our modified R2C2 protocol.

We have demonstrated that by using a modified R2C2 protocol, stringent ONT sequencing practices and Chopper, we can produce plasmid assemblies with accuracies that are comparable to commercial sequencing centers. Modifying R2C2 by eliminating the initial circularization step reduced the processing time for plasmid samples as well as the cost of sample preparation. The shortened R2C2 protocol requires one overnight incubation for RCA, one day of ONT library preparation including at least an hour of sequencing or potentially overnight, and then one day of analysis with C3POa and Chopper for a total of 3 days from sample to assembly. We further reduced the cost by decreasing the volume of key ONT library preparation reagents for each sample and negated the cost of a flow cell by sequencing on previously used flow cells without sacrificing the quality of our data.

We showed that our protocol competes with Plasmidsaurus for time and cost while still producing data of comparable quality when sequencing individual plasmids. We also demonstrated that we produce the same quality data by sequencing a pool of plasmids. Plasmidsaurus has strict guidelines for purification of plasmids and outlines quality checks to perform before submitting samples to them for sequencing. Failure to adhere to their guidelines, may result in subpar sequencing data and analysis at no fault to the company or their protocol.

Developing and testing the modified R2C2 protocol and Chopper with a pool of plasmids proves the capabilities of this protocol to handle varied input material while still maintaining quality results.

Even though the four plasmids in this study were purified by a manufacturer which also offered reference sequences for the plasmids, Chopper was developed as a de novo assembler which means it has the ability to assemble unknown plasmids as well as the potential to highlight contamination. To determine the content of plasmids without references, a researcher can verify their assembled plasmids using the plasmid annotation software pLannotate (McGuffie and Barrick 2021). The free, web-based software generates a plasmid map and identifies functional regions within the plasmids based on FASTA input. After running Chopper, researchers can upload the assembly FASTA to pLannotate and visualize their plasmids.

Methods

Library Preparation

We sourced four plasmids with known references from *addgene*: pCSCMV:tdTomato, lentiCRISPR v2, pSpCas9(BB)-2A-Puro (PAX459) V2.0, pcDNA3.1-GFP(1-10). We used 125 ng of each plasmid to create a pool before R2C2 conversion. The nature of plasmids allows us to skip the initial Gibson assembly and linear digestion steps of the R2C2 process which produces circular DNA. We started processing our plasmids with rolling circle amplification (RCA). We performed 5 reactions for each plasmid and the pool using 100 ng of input. The samples were incubated overnight with Phi29 (NEB) and random hexamer primers. The RCA product was debranched with T7 endonuclease (NEB) for 2 hours at 37°C

and then cleaned using a Monarch PCR and DNA Cleanup Kit (NEB). The purified RCA product was size-selected on an agarose gel: DNA at 10 kb and above was excised from the gel. R2C2 DNA was extracted from gel fragments using a Monarch DNA Gel Extraction Kit (NEB).

ONT libraries were prepared from R2C2 DNA using the ONT ligation sequencing kit V14 (ONT SQK-LSK114) following the manufacturer's protocol while reducing the amount of Ligation Adapter by a tenth of the recommended amount. The libraries were then sequenced on an ONT PromethION flow cell (R10.4.1) for about 1 hour on a P2 Solo. Each flow cell was previously used for other experiments and reloaded with the plasmid R2C2 DNA after nuclease flush and a report of over 100 available pores.

Analysis

Raw nanopore sequencing data in the POD5 file format were basecalled with guppy5 using ONT's "sup" setting to generate FASTQ files (dna_r10.4.1_e8.2_400bps_sup.cfg). R2C2 raw reads in FASTQ format were then processed with a software pipeline that includes C3POa (v3.2; https://github.com/christopher-vollmers/C3POa)

python3 C3POa.py -r [guppy5 FASTQ] -l [read length cutoff] -d [minimum peak distance] -ns -z

or TideHunter (v1.5.2; https://github.com/yangao07/TideHunter) (Gao et al. 2019)

TideHunter -l -m 4000 -c 2 [guppy5 FASTQ] > [consensus FASTA]

to generate full length consensus reads and Chopper (v1.0.0; https://github.com/kschimke/Chopper)

python3 Chopper.py -c [consensus FASTA] -s [subreads FASTQ]

which bins consensus reads by length and generates full length assemblies from peak bins using medaka (v1.9.1; https://github.com/nanoporetech/medaka).

medaka_consensus -m r1041_e82_400bps_sup_v4.0.0 -f -i [reads FASTA] -d [max coverage read FASTA] -o [consensus FASTA]

To determine their accuracy, the assembled plasmids were aligned to their provided references from Addgene using minimap2 (v2.18-r1015) (Li 2018).

minimap2 -ax splice --cs=long --MD --secondary=no

Conclusion

The work presented here is a demonstration of the versatility of the R2C2 protocol to improve the quality of the data produced on ONT sequencers and shorten the time to results when compared to Illumina sequencing. The R2C2 and C3POa methods are reliable and general enough to allow for a wide range of applications without significant bias. The improvements to these methods have expanded their usability and capability by increasing the base quality of ONT sequencing to be comparable with Illumina sequencing and decreasing the processing time of individual samples.

Using R2C2 and C3POa, we demonstrated that it is possible to sequence Illumina libraries on ONT instruments without jeopardizing the content of the sequencing data nor the accuracy. We developed PLNK to reduce the time between library preparation and data analysis for Illumina libraries by monitoring the content of the library during an active ONT sequencing run. We showed that within an hour of ONT sequencing, PLNK processed enough data to determine the success of the Illumina library while producing quality data.

We demonstrated the abilities of the R2C2 and C3POa methods for generating high quality single-cell data from 10X RNA libraries. Improving the base quality of ONT output with R2C2 was vital to be able to demultiplex the long reads without relying on short read data. We highlighted that the single cell data generated from R2C2 and C3POa was accurate enough to apply tools developed for Illumina data to assign cell barcodes to reads. Additionally, we were able to use the long read single cell data to investigate differential isoform usage within human lung cell types. Our analysis showed that we produced data comparable to the Illumina sequencing

data for the same samples which shows potential for using our pipeline alone for single cell analysis.

Finally, we demonstrated that modifying the R2C2 protocol and expanding the functionality of C3POa produced high quality plasmid data even with minimal sequencing parameters. We developed Chopper to assemble plasmids from consensus reads and showed that the assemblies are comparable to the data generated from commercial sequencing centers. We showed that sequencing in lab on ONT instruments is both cost effective and time efficient for validating whole plasmids.

The R2C2 and C3POa methods will continue to be enhanced and refined as new applications for these methods are explored. The marked improvement of R2C2 sequencing data over raw ONT sequencing ensures the effectiveness of the protocol and encourages lab based sequencing over commercial Illumina sequencing in terms of both quality of data and time. Our lab continues to demonstrate the versality of R2C2 and C3POa and the three applications discussed in this thesis barely scrape the surface of this protocol's limits.

References

Adams M, McBroome J, Maurer N, Pepper-Tunick E, Saremi NF, Green RE, Vollmers C, Corbett-Detig RB. 2020. One fly–one genome: chromosome-scale genome assembly of a single outbred *Drosophila melanogaster. Nucleic Acids Res* **48**: e75. doi:10.1093/nar/gkaa450

Ali SM, Hensing T, Schrock AB, Allen J, Sanford E, Gowen K, Kulkarni A, He J, Suh JH, Lipson D, et al. 2016. Comprehensive genomic profiling identifies a subset of crizotinib-responsive ALK-rearranged non-small cell lung cancer not detected by fluorescence in situ hybridization. *Oncologist* **21**: 762–770. doi:10.1634/theoncologist.2015-0497

Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Sade-Feldman M, Gatzen M, Sarkizova S, Schwartz MA, Popic V, Blaum EM, et al. 2021. High throughput RNA isoform sequencing using programmable cDNA concatenation. bioRxiv doi:10.1101/2021.10.01.462818v1

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837. doi:10.1016/j.cell.2007.05.009

Baslan T, Kovaka S, Sedlazeck FJ, Zhang Y, Wappel R, Tian S, Lowe SW, Goodwin S, Schatz MC. 2021. High resolution copy number inference in cancer using short-molecule nanopore sequencing. *Nucleic Acids Res* **49**: e124. doi:10.1093/nar/gkab812

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527. doi:10.1038/nbt.3519

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013.

Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218. doi:10.1038/nmeth.2688

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**: 1119–1125. doi:10.1038/nbt.2727

Bushnell B, Rood J, Singer E. 2017. BBMerge: accurate paired shotgun read merging via overlap. *PLoS One* 12: e0185056. doi:10.1371/journal.pone.0185056

Byrne A, Cole C, Volden R, Vollmers C. 2019. Realizing the potential of full-length transcriptome sequencing. *Phil. Trans. R. Soc. B* **374**: 201900972. doi:10.1098/rstb.2019.0097

Byrne A, Supple MA, Volden R, Laidre KL, Shapiro B, Vollmers C. 2019. Depletion of hemoglobin transcripts and long-read sequencing improves the transcriptome annotation of the polar bear (*Ursus maritimus*). *Front Genet* **10**: 643. doi:10.3389/fgene.2019.00643

Chang S, Ho D, McLaughlin JR, Chang SY. 1984. Recombination following transformation of *Escherichia coli* by heteroduplex plasmid DNA molecules. *Gene* **29**: 255-261. doi:10.1016/0378-1119(84)90054-4

Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: *de novo* genome assembly with short paired-end reads. *PLoS One* **6**: e23501. doi:10.1371/journal.pone.0023501

Cole C, Byrne A, Adams M, Volden R, Vollmers C. 2020. Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *Genome Res* **30**: 589–601. doi:10.1101/gr.257188.119

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. doi:10.1093/bioinformatics/bts635

Emiliani FE, Hsu I, McKenna A. 2022. Multiplexed Assembly and Annotation of Synthetic Biology Constructs Using Long-Read Nanopore Sequencing. *ACS Synth Biol* **11**: 2238-2246 doi:10.1021/acssynbio.2c00126

Gallegos JE, Rogers MF, Cialek CA, Peccoud J. 2020. Rapid, robust plasmid verification by de novo assembly of short sequencing reads. *Nucleic Acids Res* **48**: e106. doi:10.1093/nar/gkaa727

Gao Y, Liu B, Wang Y, Xing Y. 2019. TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* **35**: i200–i207. doi:10.1093/bioinformatics/btz376

Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**: 353–359. doi:10.1038/s41586-022-05035-y

Gohl DM, Magli A, Garbe J, Becker A, Johnson DM, Anderson S, Auch B, Billstein B, Froehling E, McDevitt SL, et al. 2019. Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification. *Genome Biol* **20**: 85. doi:10.1186/s13059-019-1691-6

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–D1186. doi:10.1093/nar/gkr944

Gu W, Deng X, Lee M, Sucu YD, Arevalo S, Stryke D, Federman S, Gopez A, Reyes K, Zorn K, et al. 2021. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat Med* **27**: 115–124. doi:10.1038/s41591-020-1105-z

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075. doi:10.1093/bioinformatics/btt086

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature* **585**: 357–362. doi:10.1038/s41586-020-2649-2

Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng* **9**: 90–95. doi:10.1109/MCSE.2007.55

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**: 1103–1110. doi:10.1038/s41592-020-00971-x

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12. doi:10.1186/gb-2004-5-2-r12

Lemon JK, Khil PP, Frank KM, Dekker JP. 2017. Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates. *J Clin Microbiology* **55**: 3530–3543, doi:10.1128/JCM.01069-17

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN]. https://arxiv.org/abs/1303.3997.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103–2110. doi:10.1093/bioinformatics/btw152

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. doi:10.1093/bioinformatics/btp352

Li C, Chng KR, Boey EJH, Ng AHQ, Wilm A, Nagarajan N. 2016. INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5**: 34. doi:10.1186/s13742-016-0140-7

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**: 10–12. doi:10.14806/ej.17.1.200

McGuffie MJ, Barrick JE. 2021. pLannotate: engineered plasmid annotation. *Nucleic Acids Res* **49**: W516–W522. doi: 10.1093/nar/gkab374

Mitchell PK, Wang L, Stanhope BJ, Cronk BD, Anderson R, Mohan S, Zhou L, Sanchez S, Bartlett P, Maddox C, et al. 2022. Multi-laboratory evaluation of the Illumina iSeq platform for whole genome sequencing of Salmonella,

Escherichia coli and *Listeria. Microb Genom* **8**: 000717. doi:10.1099/mgen.0.000717

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628. doi:10.1038/nmeth.1226

Munro R, Santos R, Payne A, Forey T, Osei S, Holmes N, Loose M. 2021. minoTour, real-time monitoring and analysis for nanopore sequencers. *Bioinformatics* **38**: 1133–1135. doi:10.1093/bioinformatics/btab780

Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE, et al. 2014. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**: 548–554. doi:10.1038/nm.3519

Nohr E, Kunder CA, Jones C, Sutton S, Fung E, Zhu H, Feng SJ, Gojenola L, Bustamante CD, Zehnder JL, et al. 2019. Development and clinical validation of a targeted RNAseq panel (Fusion-STAMP) for diagnostic and predictive gene fusion detection in solid tumors. bioRxiv doi:10.1101/870634v1

Oliphant TE. 2007. Python for scientific computing. *Comput Sci Eng* **9**: 10–20. doi:10.1109/MCSE.2007.58

Pardo-Palacios F, Reese F, Carbonell-Sala S, Diekhans M, Liang C, Wang D, Williams B, Adams M, Behera A, Lagarde J, et al. 2021. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Research Square.* https://www.researchsquare.com/article/rs-777702/v1 (accessed January 23, 2022). doi:10.21203/rs.3.rs-777702/v1

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**: 983–987. doi:10.1038/nbt.4235

Ricci WA, Levin L, Zhang X. 2020. Genome-wide profiling of histone modifications with ChIP-Seq. *Methods Mol Biol* **2072**: 101–117. doi:10.1007/978-1-4939-9865-4_9

Ruan J, Li H. 2020 “Fast and accurate long-read assembly with wtdbg2,” *Nat Methods* **17**: 55–158. doi:10.1038/s41592-019-0669-3

Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, Baid G, Kolmogorov M, Eizenga JM, Miga KH, et al. 2021. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods* **18**: 1322–1332. doi:10.1038/s41592-021-01299-w

Silvestre-Ryan J, Holmes I. 2021. Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. *Genome Biol* **22**: 38. doi:10.1186/s13059-020-02255-1

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**: 865–868. doi:10.1038/nmeth.4380

The Tabula Sapiens Consortium. 2022. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**: eabl4896. doi:10.1126/science.abl4896

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**: 396–411. doi:10.1101/gr.222976.117

Thirunavukarasu D, Cheng LY, Song P, Chen SX, Borad MJ, Kwong L, James P, Turner DJ, Zhang DY. 2021. Oncogene concatenated enriched amplicon nanopore sequencing for rapid, accurate, and affordable somatic mutation detection. *Genome Biol* **22**: 227. doi:10.1186/s13059-021-02449-1

Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, Chang S, Conley SD, Mori Y, Seita J, et al. MA. 2020. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**: 619–625. doi:10.1038/s41586-020-2922-4

Valliyodan B, Cannon SB, Bayer PE, Shu S, Brown AV, Ren L, Jenkins J, Chung CY-L, Chan T-F, Daum CG, et al. 2019. Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J* **100**: 1066–1082. doi:10.1111/tpj.14500

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**: 261–272. doi:10.1038/s41592-019-0686-2

Volden R, Vollmers C. 2022. Single-cell isoform analysis in human immune cells. *Genome Biol* **23**: 47. doi:10.1186/s13059-022-02615-z

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

Vollmers AC, Mekonen HE, Campos S, Carpenter S, Vollmers C. 2021. Generation of an isoform-level transcriptome atlas of macrophage activation. *J Biol Chem* **296**: 100784. doi:10.1016/j.jbc.2021.100784

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

Wick RR, Holt KE. 2021. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res* **8**: 2138. doi:10.12688/f1000research.21782.4

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3** doi:10.1099/mgen.0.000132

Wilson BD, Eisenstein M, Soh HT. 2019. High-fidelity nanopore sequencing of ultra-short DNA targets. *Anal Chem* **91**: 6783–6789. doi:10.1021/acs.analchem.9b00856

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, LiW, et al. 2008. Model-based Analysis of ChIPSeq (MACS). *Genome Biol* **9**: R137. doi:10.1186/gb-2008-9-9-r137