bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Illumina But With Nanopore

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

Running Title: Illumina But With Nanopore

Alexander Zee[1#], Dori Z. Q. Deng[2#], Matthew Adams[2#], Kayla D. Schimke[1#], Russell Corbett-Detig[1], Shelbi L. Russell[2], Xuan Zhang[3], Robert J. Schmitz[3], Christopher Vollmers[1*]

1) Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA
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

[#]Contributed equally to this work
*Corresponding author: vollmers@ucsc.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

High-throughput short-read sequencing has taken on a central role in research and diagnostics. Literally hundreds of different assays exist today to 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 inertia associated with the research enterprise as a whole 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 long-read sequencer, is associated with little to no capital cost. Here we show that we can make short-read Illumina libraries compatible with the long-read ONT MinION by circularizing and rolling circle amplifying the short library molecules using the R2C2 method. 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 millions of 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, as well as 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Over the last 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)[1], chromatin state (ChIP-seq[2] and ATAC-seq[3]), protein abundance[4], and of course to aid the assembly of genomes[5] - among many other things. All of these methods have in common the requirement to produce a final sequencing library that is ~200-600bp of double stranded DNA 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 last decade. However, due to the high capital cost of Illumina short-read instruments, all but the most well equipped labs outsource their Illumina sequencing to core facilities. While this provides access to the most recent sequencing technology, this outsourcing can lead to long delays between running an experiment and receiving results. Because of this, genomics assays still exist in a way that is not fully integrated in standard lab workflows. This is particularly problematic in a clinical setting when fast sample turn-around is very important or when new methods are being developed or established in a research laboratory and rapid evaluation of sequencing libraries would speed up research progress. While benchtop sequencers like the iSeq and MiniSeq by Illumina exist, they are still very expensive and even when compared to the low-throughput MiSeq come with trade-offs in throughput, accuracy, and read lengths. Other companies like Ion Torrent, Genapsys, or BGI are now offering competing sequencing instruments, but without offering a distinct advantage over Illumina's instruments[6,7]. Instrument cost and the need for modifying library preparation and often well established and understood analysis workflows for these new platforms will be met with a significant amount of inertia amongst bench and computational scientists.

Over the last 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 2x300 bp run). Intriguingly, this throughput comes with tunable read length, so the MinION can produce more shorter reads or fewer longer reads depending on how DNA is prepared for sequencing. Most importantly, ONT MinION sequencers are included in the first purchase of ONT sequencing reagents. However, standard per-base sequencing accuracy of even the newest basecalling software guppy5 is only around 96% and dominated by insertion and deletion errors which are almost absent in Illumina data. Furthermore, ONT MinIONs sequencing accuracy declines with shorter reads[8].

Here, we implemented a simple workflow that converts almost any Illumina sequencing library into DNA of lengths optimal for the ONT MinION and generated data at similar cost and accuracy as the Illumina MiSeq. We made this possible by using the previously published and optimized R2C2 method[9–14]. R2C2 circularizes dsDNA libraries, 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 and 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[13]. Since Illumina libraries are shorter than full-length cDNA, we modified the R2C2 protocol to generate a large number of shorter 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, 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. To highlight the potential of this approach for the rapid evaluation of library metrics, we developed the PLNK (Processing Live Nanopore Experiments) tool to take advantage of the real-time data generation of the ONT MinION. PLNK processes raw data and generates immediate feedback on library composition and what percentage of reads fall within defined regions in the genome. We show that this on-target percentage can be used to evaluate RNA-seq, ChIP-seq and enriched Tn5 libraries.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

To generate R2C2 data of 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 lung cancer cell lines NCI-H1650 and NCI-H1975 which we enriched for the protein coding regions of ~100 cancer relevant genes (Fig. 1).

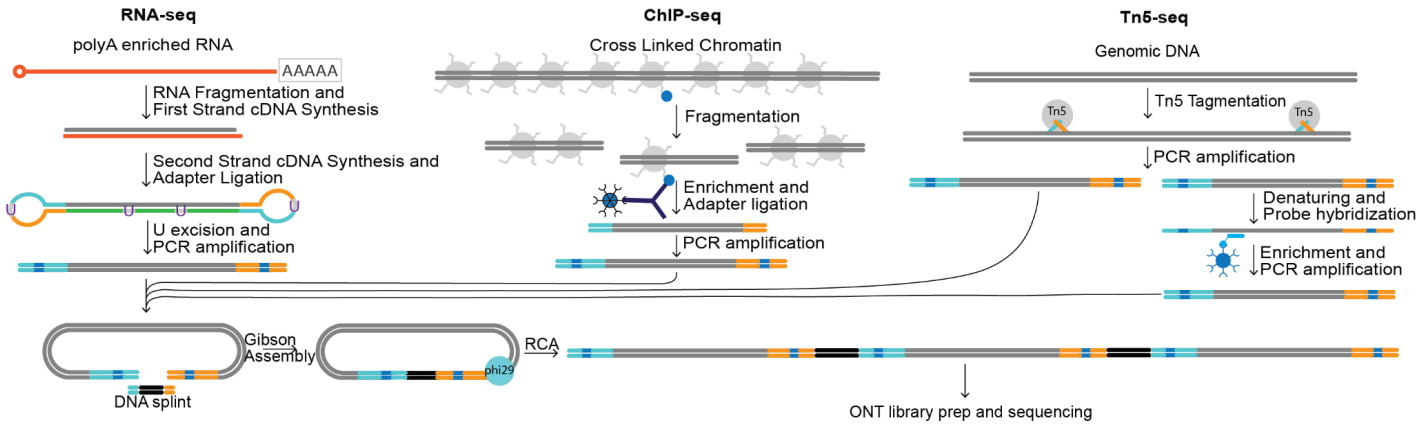

Fig. 1: Experiment overview. Illumina RNA-seq, ChIP-seq, and Tn5-based genomic libraries libraries (regular and enriched) were generated from different samples. The Illumina libraries were then circularized and amplified using rolling circle amplification (RCA). The resulting DNA, containing tandem repeats of Illumina library molecules, was then prepped for sequencing on the ONT MinION sequencer.

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 (Table S1). After rolling circle amplification of these DNA circles by Phi29 polymerase, we fragmented and size selected the 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 (Table 1). All data was then basecalled with the guppy5 *dna_r9.4.1_450bps_sup.cfg* model and consensus called using C3POa (v2.2.3) (https://github.com/rvolden/C3POa).

Library type	Organism	Raw reads (pass filter)	Raw read median length	R2C2 reads	Demuxed reads	Subreads/ R2C2 read	Per-Read accuracy
RNA-seq	Homo sapiens	9,500,956	2,288	8,992,882	8,066,704	3.14	99.59%
ChIP-seq	Glycine max (soybean)	4,518,775	3,360	4,191,438	4,023,935	3.93	99.12%
Tn5	D. melanogaster/ Wolbachia	5,188,771	2,447	3,339,161	N/A	4.88/ 4.68	98.8%/ 99.61%
Enriched Tn5	Homo sapiens	4,062,736	3,377	3,825,657	3,078,913	4.85	99.38%

Table 1: R2C2 sequencing run characteristics.

To benchmark the R2C2 data of Illumina libraries, we also sequenced the same libraries on different Illumina sequencers and compared the metrics most relevant to the different library types.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evaluating R2C2 for the sequencing of Illumina RNA-seq libraries

First, we benchmarked the ONT-based R2C2 method for the generation of RNA-seq data from Illumina libraries. To this end we sequenced Illumina RNA-seq libraries of the human lung carcinoma cell line A549 on the Illumina MiSeq and after R2C2 conversion on the ONT MinION. We prepared four technical replicates in the form of dual indexed paired-end Illumina libraries using the NEBnext Ultra II Directional RNA kit. We pooled and sequenced three of those technical replicates on a multiplexed Illumina MiSeq 2x300bp paired end run to ensure fully sequencing even the longest library molecules. We then pooled aliquots of the same replicates and converted them into R2C2 libraries and sequenced them on an ONT MinION instrument. To evaluate our ability to quantitatively pool libraries at different points in the R2C2 workflow, we processed a fourth replicate in parallel and added it at a specific ratio after rolling circle amplification.

After sequencing and consensus calling with C3POa, we compared general characteristics of the R2C2 data with the Illumina MiSeq data. The Illumina MiSeq run produced 20,830,560 2x300 bp paired-end reads with a median insert length of 237bp as determined by merging the read pairs with bbmerge[15]. The R2C2 run produced 8,992,882 R2C2 reads with a median raw read length of 2,288 bp that covered Illumina library inserts (median length 253 bp) an average of 3.14 times (Table 1). Overall the length distribution of insert lengths was very similar in R2C2 and Illumina MiSeq data (Fig. 2a).

The three technical replicate libraries in the Illumina MiSeq run were initially pooled at 4:2:1 ratio and produced 4:2.03:1.58 read ratio after demultiplexing. The same three libraries generated a similar 4:2.28:1.34 ratio in the R2C2 data after demultiplexing with differences to the Illumina MiSeq results likely being due to pipetting variability when pooling the libraries for the different sequencing methods. Further, the fourth replicate represented 40.5% of the R2C2 data which is slightly more than the 30% of R2C2 DNA it represented in the MinION sequencing run. While 9.71% of reads were not assigned to any sample, only 1.7% of reads were assigned to a combination of Illumina indexes not included in the pool. The 1.7% single index switching rate implies that theoretically only 0.0289% (1.7%*1.7%) or less than 3 out of 10,000 reads were misassigned to the wrong sample in a dual indexed library.

To compare the two technologies at the highest possible sequencing depth, we aligned and evaluated the entire non-demultiplexed datasets. We used the STAR aligner[16] (STARlong executable for R2C2 data) which is routinely used for standard Illumina RNA-seq analysis. 90.08% and 90.12% of Illumina MiSeq and R2C2 reads aligned, respectively, indicating that R2C2 reads are accurate enough to be aligned by the STAR aligner with default settings.

The resulting read alignments allowed us to use the reads' identity to the human genome to evaluate the accuracy of the Illumina MiSeq and R2C2 approaches. Based on their alignments, median read accuracy was scored as 100% for Illumina MiSeq and 99.59% for R2C2 data. However because of their short length many reads produced by either technology won't contain errors, which will distort the actual observed error rate. To evaluate accuracy independent of read length, we used the mean per-base accuracy as reported by the STAR aligner. For the Illumina MiSeq data, STAR reported per-base mismatch, deletion, and insertion rates of 0.54%, 0.01%, and 0.01%, respectively. For the R2C2 data, STAR reported per base mismatch, deletion, and insertion rates of 0.38%, 0.36%, and 0.13%, respectively. Illumina MiSeq data therefore showed a higher rate of mismatches but a lower rate of insertion and deletion than the R2C2 data. Further analysis showed in contrast to R2C2 data, Illumina base accuracy decreased with increasing read cycles, particularly in read 2, with R2C2 surpassing Illumina MiSeq accuracy for read 2 lengths over ~175 bp (Figure 2B and D).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Sequencing method	Mismatch rate per base, %	Deletion rate per base	Deletion average length	Insertion rate per base	Insertion average length
Illumina MiSeq	0.54%	0.01%	1.78	0.01%	1.97
R2C2 (ONT MinION)	0.38%	0.36%	1.50	0.13%	1.52

Table 2. RNA-seq error rates as determined by the STAR aligner

The primary output of RNA-seq experiments is the quantification of gene expression levels, typically determined by counting and normalizing the number of reads aligned to an annotated gene. A comparison of normalized gene counts as determined by STAR for Illumina MiSeq and R2C2 datasets had a Pearson's r value of 0.996 (Fig. 2C). Additionally, the identification and quantification of splice junctions can identify patterns of alternative splicing, which provides a higher resolution view of the transcriptome. Normalized splice junction counts as determined by STAR for Illumina MiSeq and R2C2 datasets had a Pearson's r 0.977 (Fig. 2C). Finally, we also tested whether ultra-fast pseudo-alignment based tools like kallisto[17] will generate reliable gene expression levels based on R2C2 reads which feature more insertion and deletion rates compared to standard Illumina data. We found that gene expression values as determined by kallisto for Illumina MiSeq and R2C2 datasets had a Pearson's r value of 0.985 (Fig. 2C).

Overall this comparison showed that using R2C2, we can convert Illumina RNA-seq libraries into DNA ideally suited for the ONT MinION. We can generate millions of reads from this DNA, highly efficiently demultiplex the resulting accurate consensus reads and use standard computational tools - STAR and kallisto - to analyze them. The gene expression values generated by R2C2 are highly similar to those generated by Illumina MiSeq data from the same libraries.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

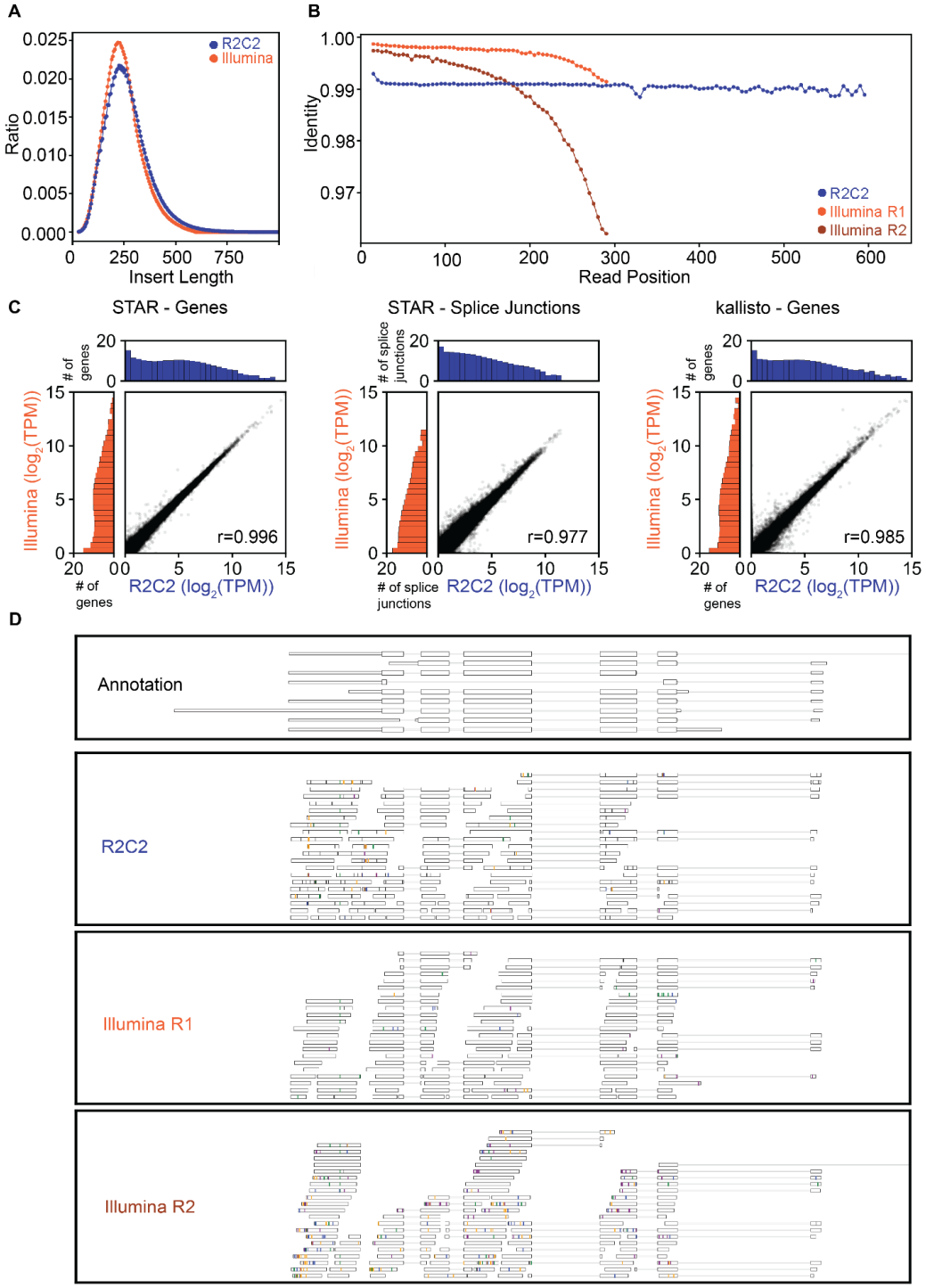

Fig. 2. Sequencing Illumina RNA-seq libraries on the ONT MinION after R2C2 conversion. Insert length distribution (A) and read position dependent identity to the reference genome (B) of R2C2 and Illumina MiSeq reads generated from the same Illumina library. C) Comparisons of R2C2 and Illumina MiSeq read-based gene expression and splice junction usage quantification by STAR and kallisto are shown as scatter plots with marginal distributions (log2 normalized) shown as histograms. D) Genome browser-style visualization of read alignments. Mismatches are marked by lines colored by the read base (A - orange; T - green; C - blue; G - purple). Insertions are shown as gaps in the alignments while deletions are shown as black lines.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evaluating R2C2 for the sequencing of Illumina ChIP-seq libraries

Next, we tested the ability of R2C2 for the quality control of Illumina ChIP-seq libraries. To do this, we converted a previously generated ChIP-seq library targeting the H3K4me3 histone modification in a *Glycine max* (soybean) sample. The H3K4me3 library and its corresponding control Input library had previously been sequenced on an Illumina NovaSeq 6000 to a depth of 8,413,865 and 32,377,813 2x150bp paired end reads, respectively. Based on their alignment, the sequenced molecule libraries had an insert length of 390 bp (H3K4me3) and 312 bp (Input).

Because the H3K4me3 and Input libraries had been prepared with only a single index distinguishing them, we converted the libraries separately with R2C2 using distinct DNA splints that contained unique index sequences. This added an extra level of indexing to minimize concerns of potential index crosstalk. We splint-indexed, and pooled H3K4me3 and Input ChIP-seq Illumina libraries and sequenced the pool on a single ONT MinION flow cell. We then demultiplexed the resulting R2C2 reads, assigning 2,493,021 and 1,530,914 reads (1.6:1) to H3K4me3 and Input libraries, respectively, which a ratio which corresponded well with the 1.35:1 ratio at which they were pooled prior to sequencing. Importantly, the demultiplexing script scored only 163,489 (3.9%) reads as “undetermined” and assigned only 4,014 (0.1%) reads to a combination of indexes not present in the library. This indicated that the extra level of indexing was highly successful in minimizing index crosstalk.

The demultiplexed R2C2 reads showed a median read accuracy of 99.23% (H3K4me3) and 98.8% (Input) as well as a median read length of 556 bp (H3K4me3) and 459 bp (Input). The molecules sequenced by R2C2 were therefore longer than the molecules sequenced by the Illumina NovaSeq 6000. The difference between the technologies is likely due to the high bias of the Illumina NovaSeq 6000 towards shorter molecules.

Sample	Illumina HiSeq Reads	Median Insert length	R2C2 Reads	Median Insert length
H3K4me3	8,413,865	389	2,493,021	556
Input	32,377,813	312	1,530,914	459

Table 3. ChIP-seq read characteristics

To test whether R2C2 reads could replace the same number of Illumina reads, we subsampled the Illumina sequencing data to the depth of the R2C2 data for both samples. We then aligned both Illumina NovaSeq 6000, subsampled Illumina NovaSeq 6000, and R2C2 reads to the *Glycine max* genome (Gmax_508_v4.0)[18]. For alignment, we chose the short-read preset of the minimap2[19] aligner for both Illumina and R2C2 data. We then called peaks on the full H3K4me3 Illumina NovaSeq 6000 dataset using MACS2 and tested whether both subsampled Illumina NovaSeq 6000 and R2C2 data could be used to evaluate the success of a ChIP experiment. Visual inspection of the data using the Phytozome JBrowse genome browser[20] as well as our own tools (Fig. 3D) showed that subsampled Illumina NovaSeq 6000 and R2C2 data both demonstrate the same enrichment patterns as the full Illumina NovaSeq 6000 data. A systematic analysis showed that 84% of R2C2 reads and 69% of subsampled Illlumina reads overlap with an H3K4me3 peak identified on the full Illumina data, whereas only 18% and 11% of Input reads do so.

To compare whether R2C2 and subsampled Illumina NovaSeq 6000 datasets are also similar quantitatively, we counted how many reads for each of the datasets fell into each H3K4me3 peak we identified using the full Illumina NovaSeq 6000 dataset and MACS2. We found that the peak depths are correlated (Pearson’s r=0.776). This correlation is lower than we observed with the RNA-seq data, likely because the ChIP-seq library had a longer insert size than the RNA-seq library. Because the Illumina NovaSeq 6000 preferentially amplifies and sequences shorter molecules, it ultimately distorts the library composition and sequences a different set of molecules than the R2C2 approach. This means that while R2C2 can be used to evaluate whether a ChIP-seq experiment successfully enriched chromatin from expected genomic regions, R2C2 will not produce quantification highly similar to the Illumina NovaSeq 6000 because R2C2 data more closely reflects the actual composition of the sequenced library.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

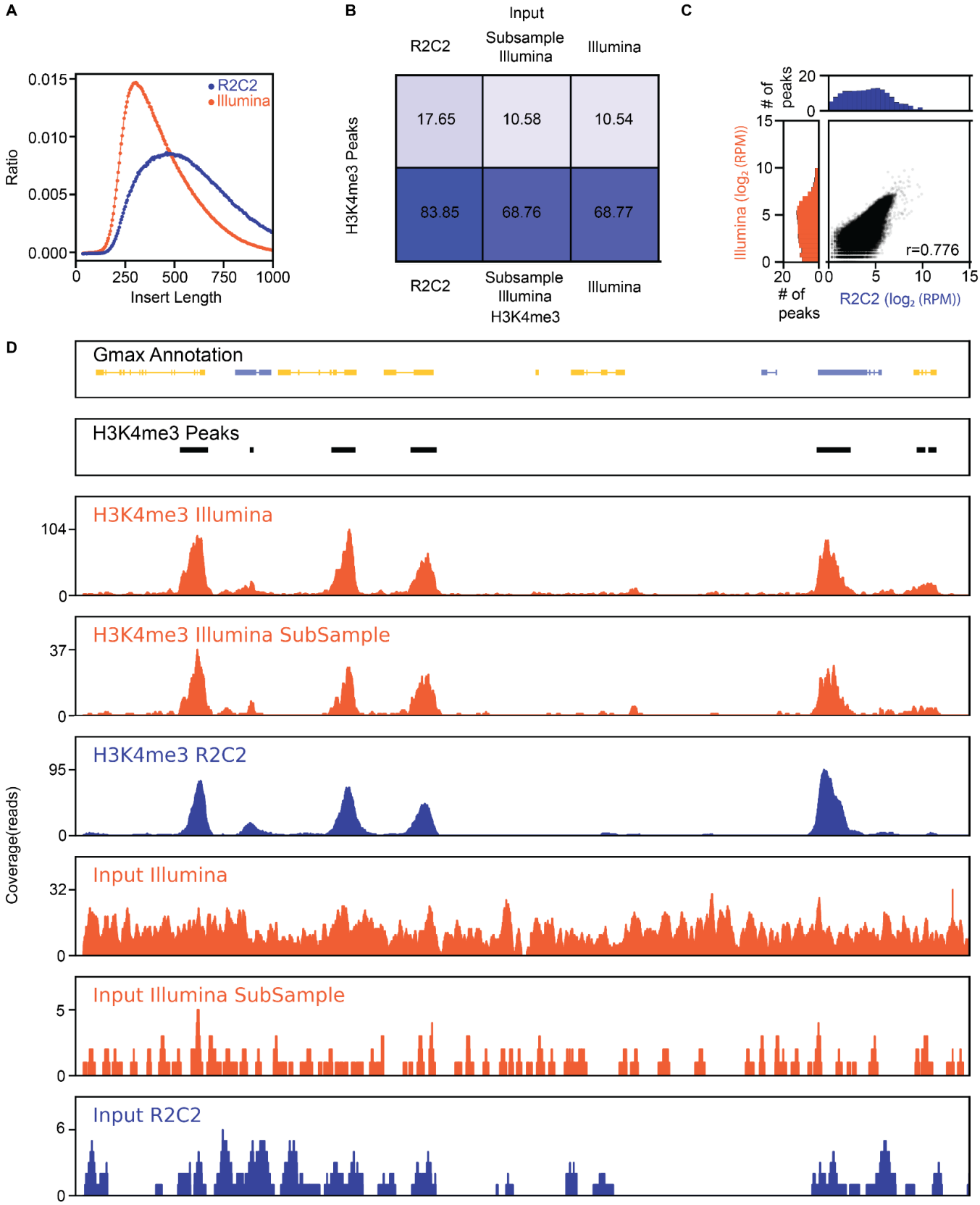

Fig. 3. Sequencing Chip-seq libraries on the ONT MinION after R2C2 conversion. A) Insert length distribution of R2C2 and Illumina NovaSeq 6000 reads generated from the same Illumina library. B) Percentage of reads in the R2C2, Subsampled Illumina and full Illumina datasets overlapping with H3K4me3 peaks generated from the full Illumina H3K4me3 dataset using MACS2. C) Comparison of the number of R2C2 and subsampled Illumina reads overlapping with H3K4me3 peaks is shown as scatter plots with marginal distributions shown as histograms. Pearson's r is shown in the bottom right. D) Genome annotation, H3K4me3 peak areas and read coverage histograms are shown for a random range in the Gmax genome.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evaluating R2C2 for the sequencing of Illumina Tn5 libraries

We tested whether R2C2 could be used to assemble small genomes from Illumina libraries generated using Tn5-based tagmentation. For this test, we chose the 1.2 Mb genome of the *Wolbachia* bacterial endosymbiont of *Drosophila melanogaster* and prepared Tn5 libraries from DNA extracted from *Wolbachia*-containing *Drosophila melanogaster* S2 cells.

While longer molecules are beneficial for genome assembly, Illumina sequencers struggle to generate clusters and sequence molecules longer than 600 bp. However, Tn5 generates molecules longer than this limit and R2C2 can efficiently process and sequence them. Therefore, instead of R2C2 converting the exact same library we used for Illumina sequencing, we chose to size-select an Illumina library. We size-selected a Tn5 library for molecules between 800-1200 bp lengths, corresponding to genomic DNA inserts of ~600-1000 bp. We then R2C2 converted and sequenced this size-selected Illumina library on the ONT MinION. We generated a total of 3,338,280 R2C2 consensus reads with a median length of 680 bp. Out of these reads, we assembled 879,303 reads that did not align to the *Drosophila melanogaster* genome. We used miniasm[21] for this assembly task and polished the resulting assembly using Medaka (v.1.4.4; https://github.com/nanoporetech/medaka). The resulting assembly contained 95 contigs which covered 97.2% of the *Wolbachia* genome, had a NGA50 of 29,963 bp and 8.5/5.6 mismatches/indels per 100 kb of sequence.

We also generated an assembly from Illumina Nextseq 2x150 bp generated from a non-size selected Tn5 library of the same cell line. From 2,552,018 2x150 bp Illumina reads we extracted 779,206 reads that did not align to the *Drosophila melanogaster* genome and assembled those reads using Meraculous[22]. The resulting assembly contained 136 contigs which covered 91.6% of the *Wolbachia* genome, had a NGA50 of 23,217 bp and 0.5/0.6 mismatches/indels per 100 kb of sequence. Neither assembly had misassemblies as determined by QUAST[23].

Comparing Illumina and R2C2 assemblies of the *Wolbachia* genome (NC_002978.6) showed R2C2 can generate more contiguous and complete assemblies from the same library type. However, systematic errors produced by the ONT MinION cannot be fully removed by the R2C2 consensus process or medaka polishing. The assembly we generate does therefore have more mismatches and indel errors than its Illumina counterpart. This ultimately suggests that when limited to a single Tn5 library due to sample constraints, R2C2 can be a valuable addition to an assembly effort but that depending on use case, further polishing with Illumina data might be required to achieve the desired base accuracy.

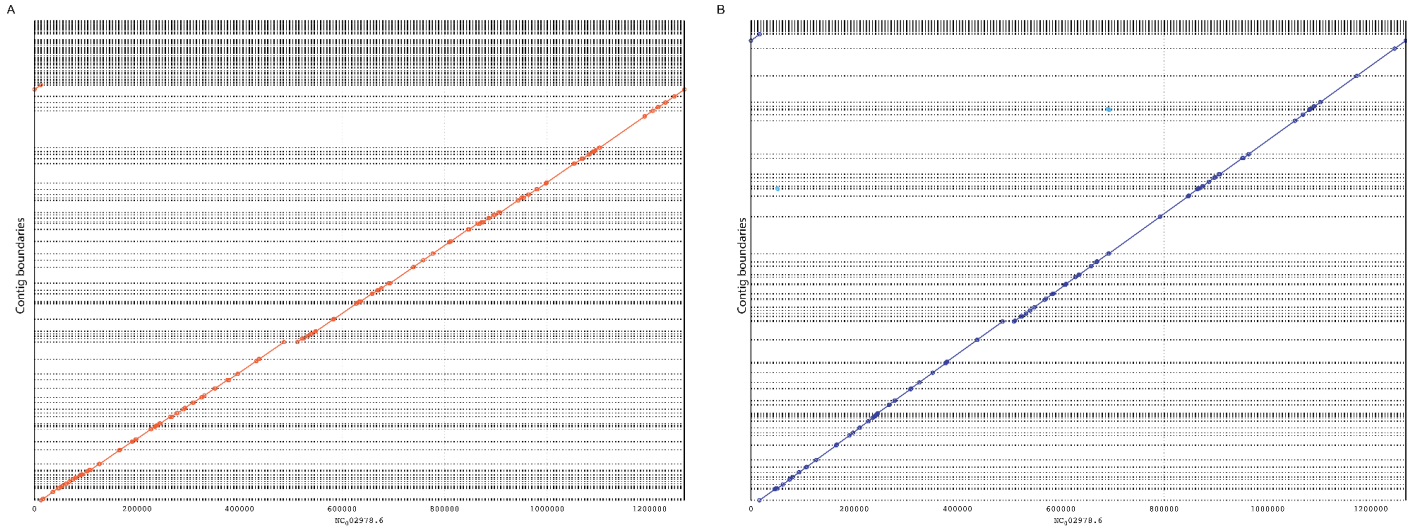

Fig. 4 Comparing R2C2 and Illumina based assemblies of a small genome. Illumina 2x150 reads were assembled in 134 contigs using Meraculous. R2C2 reads were assembled using Miniasm into 95 contigs which were polished using Medaka. The alignments of the contigs of both assemblies - (A) Illumina and (B) R2C2 - are shown as dotplots generated by mummer[24]. Both approaches fail to assemble a section of the Wolbachia genome that contains pseudogenes and a transposable element near to coordinate 500,000.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evaluating R2C2 for the sequencing of target-enriched Illumina Tn5 libraries

We tested the ability of R2C2 to evaluate target-enriched Tn5 libraries and benchmark our ability to detect germline variants in the resulting data. To this end, we generated dual-indexed Tn5 libraries from genomic DNA of two cancer cell lines (NCI-H1650 and NCI-H1975) with known mutations in the EGFR gene. We pooled these libraries and enriched the pool for a panel of cancer genes based on the Stanford solid tumor STAMP panel[25] using a Twist Bioscience oligos panel and reagents (Table S2). We performed this enrichment experiment once and without optimization. To compare R2C2 and Illumina MiSeq, we sequenced these enriched Tn5 libraries on 1) a multiplexed Illumina MiSeq 2x300 bp paired end run and 2) on an ONT MinION after R2C2 conversion.

The multiplexed MiSeq run generated 7,430,624 read pairs for the NCI-H1650 library and 1,142,187 read pairs for the NCI-H1975 library. The ONT MinION run generated 3,825,657 R2C2 reads after C3POa processing. Demultiplexing then assigned 2,057,155 (53.7%) R2C2 reads to the NCI-H1650 library and 1,021,758 (26.7%) R2C2 reads to NCI-H1975. Although 537,997 (14.1%) reads were not assigned to any sample, only 5.4% of reads were assigned to one of the two combinations of Illumina indexes not included in the pool. The 5.4% single index switching rate this implies that assuming index switches occur independently only 0.29% (5.4%*5.4%) or 3 out of 1,000 reads would be misassigned to the wrong sample in our dual indexed library.

After demultiplexing we compared the insert length and target enrichment across samples and methods. We did so by merging the Illumina MiSeq read pairs using bbmerge. As with the ChIP-seq experiment, R2C2 data showed longer insert lengths than the Illumina MiSeq, with the R2C2 insert length more closely resembling the actual length of the input library (Fig. 5A, D, and S1). We aligned the reads of different samples and methods to the human genome using the short-read preset of minimap2 and determined what percentage of reads overlapped with a target region and what coverage that amounted to in each condition. For NCI-H1650, 15.8% of R2C2 reads and 14.4% of Illumina MiSeq reads overlapped with a target region amounting to a median coverage of 128 (5th percentile: 28; 95th percentile: 310) for R2C2 and 558 (5th percentile: 134; 95th percentile: 1220) for Illumina MiSeq. For NCI-H1975, 18.5% of R2C2 reads and 16.8% of Illumina MiSeq reads overlapped with a target region amounting to a median coverage of 69 (5th percentile: 13; 95th percentile: 166) for R2C2 and 110 (5th percentile: 23; 95th percentile: 225) for Illumina MiSeq. The per-base coverage of R2C2 and Illumina MiSeq datasets was very well correlated within samples with NCI-H1650 showing a Pearson's r=0.91 and NCI-H1975 showing a Pearson's r=0.89 (Fig. 5B and E).

Next, we used the read alignments to determine per-base accuracy levels for all samples and method combinations. The NCI-H1975 sample - which also produced less reads than expected on the Illumina MiSeq - produced reads at lower than expected accuracy. Read alignments suggested that the average per-base accuracy for read 1 and read 2 in NCI-1975 were 96.81% and 98.26% compared to 98.37% and 97.88% for NCI-H1650. As expected the per-base accuracy was highly position dependent and declined with increasing sequencing cycle number (Fig. 5C and F). Further, the actual accuracy of the MiSeq reads is likely even lower due to alignments not being extended once read and genome are too dissimilar. The accuracy of R2C2 reads in both NCI-H1975 and NCI-H1650 were similar and stable throughout the reads at 98.40% and 98.28%, meaning that, in this case, the R2C2 reads had a higher per-base accuracy than the combined MiSeq reads.

Visualizing Illumina MiSeq and the R2C2 read alignments showed that both methods successfully enriched for (Fig 5G) and detected the 15 base pairs heterozygous deletion in the EGFR gene in the NCI-H1650 cell line and the C to T heterozygous variants in the EGFR gene in the NCI-H1975 cell line (Fig. 5H). To systematically evaluate the germline variant detection ability of Illumina MiSeq and R2C2 reads, we used Deepvariant[26] for calling germline variants based on the Illumina MiSeq data and Pepper-DeepVariant[27], a variant caller designed for nanopore datasets, for calling germline variants in the R2C2 sequencing results. Because of the poor sequencing performance of the Illumina MiSeq for the NCI-H1975 library, we only performed this analysis on NCI-H1650. For NCI-H1650, Illumina/Deepvariant detected 119 variants in the enriched genomic regions

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

when using a QUAL cut-off of >=33.3. R2C2/Pepper-Deepvariant detected 122 variants in the enriched genomic regions when using a QUAL score >= 3.8 including 117 of the 119 Illumina/Deepvariant calls. When we used Illumina/Deepvariant variants as ground truth, the R2C2/Pepper-Deepvariant method achieved 98.3% recall and 95.9% precision.

Overall this showed that R2C2 can accurately quantify what percentage of molecules in an enriched Tn5 Illumina library overlap with a target region. Despite showing longer insert lengths than the Illumina MiSeq dataset, the R2C2 dataset showed per-base coverage that was highly correlated with the Illumina MiSeq data. Interestingly in this experiment, R2C2 actually showed a higher per-base accuracy than the Illumina MiSeq. However, due to the remaining error in the R2C2 data likely not being random, variants called based on Illumina MiSeq and R2C2 data are very similar yet not identical, with R2C2 data showing some likely false positive calls. This highlights the persisting limitation of even error-corrected ONT data where increasing per-base accuracy is of limited utility if the remaining errors are systematic in nature.

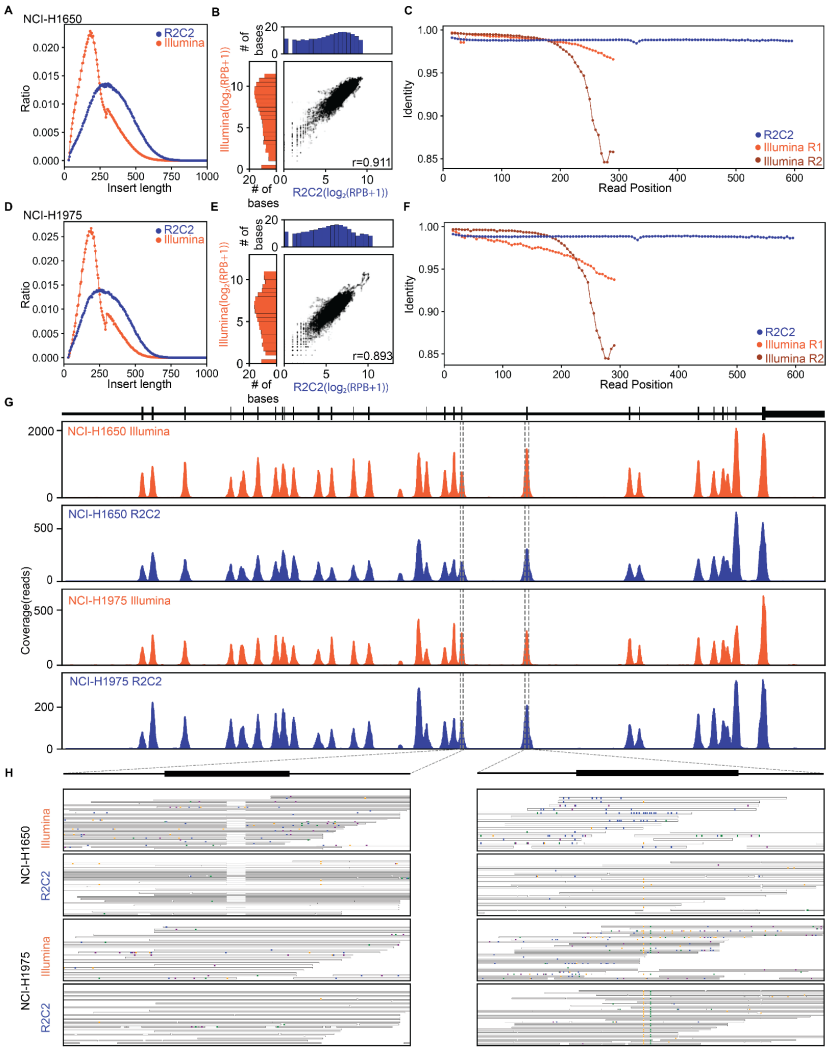

Fig. 5 Evaluating target-enriched Tn5 libraries with R2C2. A and D) Inserts length of library molecules sequenced by Illumina or R2C2 approaches. B and E) Comparison of per-base coverage in Illumina and R2C2 datasets. Marginal distributions are log2 normalized. C and F) Alignment based read position dependent accuracy shown for the indicated sequencing reads and methods. Sequencing coverage plot of the target-enriched Tn5 libraries for R2C2 and Illumina results at chromosome 7:55,134,584-55,211,629 which covers a part of the EGFR gene. Top panel shows the annotation of one EGFR isoform. The x axis of the coverage plot is the base pair position and the y axis is the total number of reads at each position. The dotted lines indicate zoomed-in views of exons that contain the 15 bps deletion in NCI-H1650 (left) and the C to T and T to G point mutations in NCI-H1975 (right). Both samples' Illumina reads and the R2C2 read alignments of the selected regions are shown. The mismatches are colored based on the read base (A - orange; T - green; C - blue; G - purple).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Real-Time Analysis of Illumina library metrics using PLNK

To enable the rapid evaluation of Illumina sequencing library metrics by R2C2, we created the computational pipeline PLNK (Processing Live Nanopore Experiments). 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, 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 processes fast5 files individually in the order they are generated. To do this, PLNK controls several external tools: guppy5 for basecalling, C3POa for R2C2 consensus generation, a separate python script for demultiplexing (based on splint sequences and Illumina indexes), and mappy (minimap2 python library) for aligning reads to a provided genome (Fig. 6A).

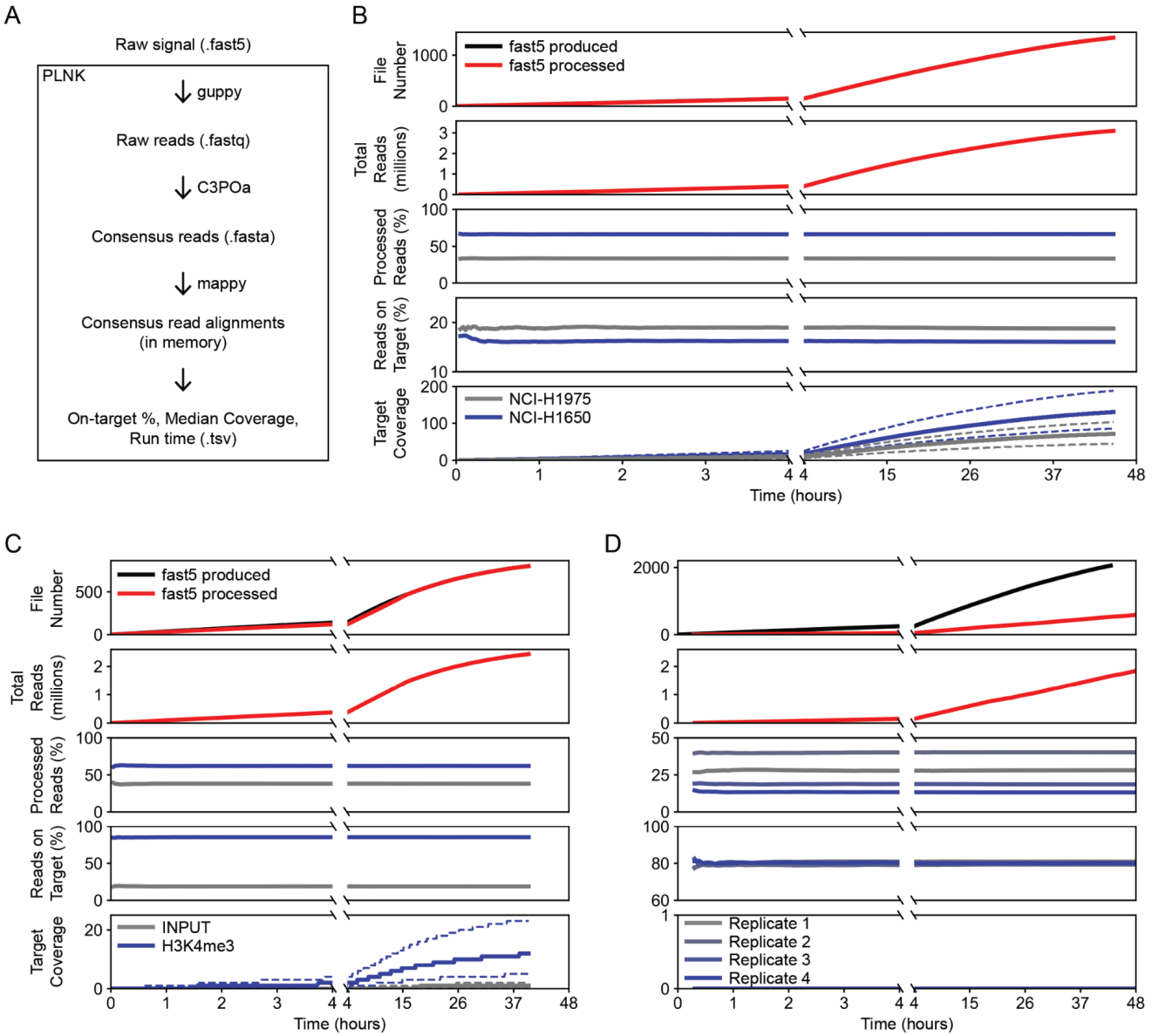

Fig. 6: Real-time characterization of Illumina sequencing libraries. A) Diagram of PLNK functionality, fast5 files processed in the order they are produced. PLNK controls guppy5 for basecalling, C3POa for consensus calling, mappy for alignment, and calculates metrics based on those alignments. B-D) Simulation of real-time analysis for enriched Tn5 (B), ChIP-seq (C), and RNA-seq (D) libraries. For each timepoint, panels from top to bottom show 1) The number of fast5 files 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 percent of reads overlapping with target regions 5) The median read coverage of bases in the target regions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 previously completed sequencing experiments. We used the metadata of fast5 files 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 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. 6B).
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. Since the RNA-seq experiment produced the largest amount of data in the study, 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 after which a ONT MinION flow cell can be flushed, stored, and reused.

Overall this suggests that PLNK can be used to evaluate Illumina library metrics in real-time. The bottleneck for analysis in our desktop computer setup seemed to be the guppy5-based basecalling using the slower yet most accurate “sup” basecalling setting. While 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.

Discussion

The capabilities of the dominant Illumina sequencing technology - producing massive numbers of very short reads at moderate accuracy - have shaped the development of sequencing based assays more than any other single factor. While long-read sequencers by Pacific Biosciences (PacBio) and ONT have now clearly superseded Illumina instruments as the gold standard technology for genome assembly, there are still hundreds of assays adapted for very short Illumina 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 work by either concatenating [OCEAN, MAS-Iso-Seq][8,28] or otherwise preparing[29] short molecules for sequencing on the PacBio or ONT instrument. While these assays can generate more short reads, they either have to contend with the high cost of the PacBio Sequel IIe sequencer, or the low per-base accuracy of raw ONT reads which even with the latest guppy5 algorithm is only 96% in our hands.

Taking inspiration from the highly accurate but throughput-limited PacBio IsoSeq and HiFi workflows, circularizing-based [R2C2[12], INC-seq[30], HiFRe[31]] methods have been developed to trade throughput for accuracy on ONT MinION and PromethION sequencers. Using a modified R2C2 method we present here, we

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 takes full advantage of the ONT MinION's throughput. Overall, these R2C2 libraries produce data with sequencing accuracy comparable to an Illumina MiSeq 2x300 bp run but do so in a read-position independent manner. By generating up to 8.99 million reads (8.1 million demultiplexed) from a single ONT MinION flow cell, this approach can even be cost competitive with the Illumina MiSeq - even without taking instrument cost into account.

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 was 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. The notable exceptions 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[32]. The only true limitation of the R2C2 method is rooted in the systematic error of the ONT sequencing platform that even the R2C2 consensus read approach cannot correct. As a result, even machine learning based polishing (Medaka) or variant calling (Pepper-Deepvariant) tools used in this study couldn't quite achieve Illumina-style performance for assembly consensus accuracy or germline variant calling, respectively. In either case, it certainly didn't help that both Medaka and Pepper-Deepvariant had been trained on raw ONT data, not R2C2 consensus reads derived from raw ONT data. Improved consensus tools[33] and consistently improving ONT sequencing accuracy will no doubt also improve R2C2-based performance for these types of analysis in the future.

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[34] or with more powerful tools like MinoTour[35]. The PLNK tool we developed here controls basecalling by the guppy5 basecaller, C3POa processing, mappy-based alignment and on-target estimation for enrichment analysis. Using this script, we showed that key metrics of RNA-seq, ChIP-seq and enriched Tn5 libraries can be evaluated in under 1 hour of sequencing. This can accelerate quality control and enables the reuse of MinION flowcells, thereby reducing sequencing cost.

In summary, we have shown that, using R2C2, the ONT MinION can - with few 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 directly and immediately after a library is produced thereby moving genomics assay from sequencing core facilities back into the lab. Even if sequencing libraries are ultimately sequenced in a sequencing core facility on an Illumina HiSeq or NovaSeq 6000 to take advantage of the extremely high throughput these instruments provide, R2C2 can be used to rapidly evaluate library pool compositions and metrics before committing to the cost and turnaround time this requires.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Library Preparation

RNA-seq

Four RNAseq 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 polyA selected RNA from the human lung carcinoma cell line A549 (Takara #636141) was used as input. The RNA fragmentation step was performed at 94C for 5 minutes. PCR enrichment of adaptor ligated DNA was performed for 9 cycles using the NEBNext Multiplex Oligos for Illumina (NEB #E7600S) kit to add Illumina dual index sequences. Three libraries were pooled at a 4ng, 2ng, and 1ng before sequencing on an Illumina MiSeq instrument for paired end 2x300 bp sequencing. The same three RNAseq libraries were pooled again at the same ratio for further R2C2 library preparation. For the R2C2 run, the fourth RNA-seq library was prepared and R2C2 converted independently and added right before ONT library preparation.

ChIP-seq

Chromatin immunoprecipitation (ChIP) was performed following the detailed protocol of Ricci et al. with minor modification[36]. 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, 25µl Dynabeads Protein A (Thermo Fisher Scientific, 10002D) were washed with ChIP dilution buffer and then incubated with 2µg 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 were transferred to new tubes. At this point, 10 µl of ChIP input aliquots were collected. Sonicated chromatin was diluted tenfold in ChIP dilution buffer to bring the SDS buffer concentration down to 0.1%. The diluted chromatin was incubated with antibody-coated beads at 4 °C overnight, then washed and reverse-crosslinked. The library was prepared in accordance with the referenced protocol.

Tn5

Genomic DNA from a *Wolbachia*-containing *Drosophila Melanogaster* cell line was extracted using a lysis-buffer plus SPRI-bead purification. The Tn5 reaction was then performed using 1ul (22ng) of this genomic DNA, 1ul of the loaded Tn5-AR, 1ul of the loaded Tn5-BR, 13 ul of H2O and 4 ul of 5× TAPS-PEG buffer and incubated at 55°C for 8 minutes (Table S1). The Tn5 reaction was inactivated by cooling down to 4°C and the addition of 5 µl of 0.2% sodium dodecyl sulphate then incubated for 10 minutes. 5 ul of the resulting product was nick-translated at 72°C for 5 minutes and further amplified using KAPA Hifi Polymerase (KAPA) using Nextera Index primers with an incubation of 98°C for 30 s, followed by 16 cycles of (98°C for 20 s, 65°C for 15 s, 72°C for 30s) with a final extension at 72°C for 5 min. Before R2C2 conversion, the resulting Tn5 library was size-selected for molecules between 800-1200bp on a 1% low-melt agarose gel.

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 100ng 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 1ul of the gDNA, 1ul

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of the loaded Tn5-AR, 1ul of the loaded Tn5-BR, 13 ul of H2O and 4 ul of 5× TAPS-PEG buffer and incubated at 55°C for 8 minutes. The Tn5 reaction was inactivated by cooling down to 4°C and the addition of 5 µl of 0.2% sodium dodecyl sulphate then incubated for 10 minutes. 5 ul of the resulting product was nick-translated at 72°C for 5 minutes and further amplified using KAPA Hifi Polymerase (KAPA) using Nextera_Primer_B_Universal and Nextera_Primer_A_Universal (Smart-seq2) with an incubation of 98°C for 30 s, followed by 16 cycles of (98°C for 20 s, 65°C for 15 s, 72°C for 30s) with a final extension at 72°C for 5 min. 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 294ng of NCI-H1975 Tn5 library, 360ng of NCI-H1650 Tn5 library, 8ul of blocking oligo pool [100uM], 8ul of universal blockers, 5ul of blocker solution and 4ul of the custom panel. The mix was dehydrated using SpeedVac and was resuspended in 20ul Fast Hybridization mix at 65C. After the addition of 30 ul of Hybridization Enhancer, the mixture was incubated at 95C for 5 minutes and 60C for 4 hours. After hybridization, the reaction mix was incubated with pre-washed 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 was used directly for reamplification with Universal primers and Equinox Library Amp Mix. The mixture was incubated at 98°C for 45 s, followed by 16 cycles of (98°C for 15 s, 65°C for 30 s, 72°C for 30s) with a final extension at 72°C for 1 min. 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 (Table S1). 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 hours at 37C then cleaned using a Zymo DNA Clean & Concentrator column-5 (Zymo #D4013). The cleaned RCA product was digested using NEBNext dsDNA Fragmentase (NEB #M0348) following the manufacturer protocol with a 10 minute 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 2-10 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 was prepared for nanopore sequencing using the ONT ligation sequencing kit (ONT #SQK-LSK110) following the manufacturer's protocol then sequenced on an ONT MinION flow cell (R9.4.1). Additional library was loaded on the same flow cell after nuclease flush.

Illumina sequencing

Library pools were sequenced either on the Illumina MiSeq using 2x300 (RNA-seq and target enriched Tn5 libraries), the Illumina NextSeq500 2x150 (Tn5 library) or the Illumina HiSeq (ChIP-seq)

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Analysis

R2C2

Raw nanopore sequencing data in the fast5 file format was basecalled using the "sup" setting of guppy5 to generate fastq files. The raw reads in fastq format were then processed by C3POa (v.2.2.3 - https://github.com/rvolden/C3POa) to generate accurate consensus reads. For C3POa postprocessing, the --trim setting and the following adapter sequences were used

```
>3Prime_adapter
CAAGCAGAAGACGGCATACG
>5Prime_adapter
AATGATACGGCGACCACCGAGATCT
```

Custom scripts (available at https://github.com/kschimke/PLNK) were used to demultiplex R2C2 consensus reads based on the sequences of their DNA splints and Illumina indexes and to trim the rest of the Illumina sequencing adapters.

RNA-seq

Illumina reads were adapter trimmed using cutadapt (v3.2)[37]

```
cutadapt -m 30 -j 50 -a AGATCGGAAGAGC -A AGATCGGAAGAGC
```

Illumina and R2C2 reads were aligned to the human genome (hg38) using STAR and STARlong (v2.7.3a)[16]

```
STAR --quantMode GeneCounts --outSAMattributes NH HI NM MD AS nM jM jI XS
```

To determine insert length, Illumina read pairs were merged using bbmerge (v38.92) with default settings.

ChIP-seq

Illumina reads were sub-sampled using a custom script (https://github.com/alexanderkzee/BWN) to match the total reads from the corresponding R2C2 library.

Illumina and R2C2 reads were aligned to the *Glycine Max* genome (Gmax_508_v4.0) using minimap2 (v2.18-r1015)[19].

```
minimap2 -ax sr --cs=long --MD
```

Peaks in H3K4me3 Illumina data were called using MACS2[38]

```
macs2 callpeak -t K4.bam -c INPUT.bam -f BAM -n K4_Illumina --nomodel --extsize 200
```

Tn5

R2C2 reads were aligned to the *Drosophila melanogaster* genome (dm6) using minimap2 ((v2.18-r1015)

```
minimap2 -ax sr --cs=long --MD
```

R2C2 reads that didn't align to the Drosophila genome were then assembled using miniasm

```
minimap2 -x ava-ont [dehosted r2c2 file] [dehosted r2c2 file] > [ava paf file]
miniasm -f [dehosted r2c2 file] [ava paf file] -m 450 -s 250 > [gfa raw assembly]
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We aligned Illumina reads to the Drosophila melanogaster genome (dm6) using bwa mem[39] under default parameters. We then extracted the sample IDs for reads that did not map to the host genome and extract that set from the raw fastq files.

Illumina reads that didn't align to the Drosophila genome were then assembled using meraculous setting the minimum contig depth to 10, expected genome size to 0.013, and using a k-mer of 51 and otherwise default parameters.

Target-enriched Tn5

Illumina reads were adapter trimmed using cutadapt (v3.2)

```
cutadapt -m 30 -j 50 -a AGATCGGAAGAGC -A AGATCGGAAGAGC
```

Trimmed Illumina and R2C2 reads were aligned to the human genome (hg38) using minimap2 (v2.18-r1015).

```
minimap2 -ax sr --cs=long --MD
```

Germline variants in Illumina data of NCI-H1650 were called using Deepvariant[26]. Germline variant in R2C2 data of NCI-H1650 were called using Pepper-Deepvariant[27]

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 hours 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.

samtools[40] (v1.11-18-gc17e914) was used extensively during analysis for sam file processing. Python[41], matplotlib[42], numpy[43], and scipy[44] were all used to analyze and visualize the data

Data availability

All data generated and analyzed for this paper are available as Bioproject PRJNA775962 at the SRA.

Code availability

All code used for analysis is available at the following github repositories
https://github.com/kschimke/PLNK
https://github.com/alexanderkzee/BWN
https://github.com/rvolden/C3POa
as indicated throughout the method section.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We thank the UCSC Paleogenomics Lab sequencing facility for sequencing RNA-seq, Tn5 and Enriched Tn5 libraries. We also want to thank Kishwar Shafin for his support with running Pepper-Deepvariant. We acknowledge funding by the National Institute of General Medical Sciences / National Institutes of Health Grant R35GM133569 (to C. V.) and R35GM128932 (to R.C.-D.). This study was funded with support from the NSF (IOS-1856627) and the United Soybean Board to R.J.S..

Author contributions

A.Z. led the R2C2 conversion of ChIP-seq libraries and Tn5 libraries, analyzed ChIP-seq data, wrote and edited the manuscript.
D.D. led the generation and R2C2 conversion of target-enriched Tn5 libraries, analyzed the target-enriched Tn5 data, wrote and edited the manuscript.
M.A. optimized the R2C2 method for the conversion of Illumina libraries, led the generation and R2C2 conversion of RNA-seq libraries, analyzed the RNA-seq data, wrote and edited the manuscript.
K.S. led the development of the PLNK real-time analysis tool, performed real-time analysis simulations, wrote and edited the manuscript.
R.C.D. conceptualized and supervised the generation of Tn5 libraries, analyzed the Tn5 library based genome assemblies, and edited the manuscript.
S.R. conceptualized the generation of Tn5 libraries, generated *Wolbachia*-containing *Drosophila melanogaster* cell lines, and edited the manuscript.
X.Z. generated and supervised the Illumina sequencing of ChIP-seq libraries, and edited the manuscript.
R.J.S. designed and supervised the generation of ChIP-seq libraries and data, and edited the manuscript.
C.V. conceptualized and designed the project, supervised the generation of data, supervised Illumina and R2C2 data analysis, performed germline variant calling on target-enriched Tn5 Illumina and R2C2 data, wrote and edited the manuscript

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
2. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
4. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
5. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
6. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
7. Foox, J. *et al.* Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat. Biotechnol.* **39**, 1129–1140 (2021).
8. Thirunavukarasu, D. *et al.* Oncogene Concatenated Enriched Amplicon Nanopore Sequencing for rapid, accurate, and affordable somatic mutation detection. *Genome Biol.* **22**, 227 (2021).
9. Volden, R. & Vollmers, C. Highly Multiplexed Single-Cell Full-Length cDNA Sequencing of human immune cells with 10X Genomics and R2C2. *Cold Spring Harbor Laboratory* 2020.01.10.902361 (2020) doi:10.1101/2020.01.10.902361.
10. Cole, C., Byrne, A., Adams, M., Volden, R. & Vollmers, C. Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *Genome Res.* **30**, 589–601 (2020).
11. Byrne, A. *et al.* Depletion of Hemoglobin Transcripts and Long-Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (Ursus maritimus). *Front. Genet.* **10**, 643 (2019).
12. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* (2018)

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

doi:10.1073/pnas.1806447115.

13. Vollmers, A. C., Mekonen, H. E., Campos, S., Carpenter, S. & Vollmers, C. Generation of an Isoform-level transcriptome Atlas of Macrophage Activation. *Journal of Biological Chemistry* 100784 (2021) doi:10.1016/j.jbc.2021.100784.
14. Adams, M. *et al.* One fly--one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster. *Nucleic Acids Res.* **48**, e75–e75 (2020).
15. Bushnell, B., Rood, J. & Singer, E. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
16. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
17. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
18. Valliyodan, B., Cannon, S. B., Bayer, P. E. & Shu, S. Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant* (2019).
19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
20. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).
21. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
22. Chapman, J. A. *et al.* Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* **6**, e23501 (2011).
23. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
24. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
25. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
26. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
27. Shafin, K. *et al.* Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

neural networks. *bioRxiv* 2021.03.04.433952 (2021) doi:10.1101/2021.03.04.433952.

28. Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmable cDNA concatenation. *bioRxiv* 2021.10.01.462818 (2021) doi:10.1101/2021.10.01.462818.
29. Baslan, T. *et al.* High resolution copy number inference in cancer using short-molecule nanopore sequencing. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab812.
30. Li, C. *et al.* INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5**, 34 (2016).
31. Wilson, B. D., Eisenstein, M. & Soh, H. T. High-Fidelity Nanopore Sequencing of Ultra-Short DNA Targets. *Anal. Chem.* **91**, 6783–6789 (2019).
32. Gohl, D. M. *et al.* Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification. *Genome Biol.* **20**, 85 (2019).
33. Silvestre-Ryan, J. & Holmes, I. Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. *Genome Biol.* **22**, 38 (2021).
34. Gu, W. *et al.* Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* **27**, 115–124 (2021).
35. Munro, R. *et al.* MinoTour, real-time monitoring and analysis for Nanopore Sequencers. *bioRxiv* 2021.09.10.459783 (2021) doi:10.1101/2021.09.10.459783.
36. Ricci, W. A., Levin, L. & Zhang, X. Genome-Wide Profiling of Histone Modifications with ChIP-Seq. *Methods Mol. Biol.* **2072**, 101–117 (2020).
37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
38. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
39. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303. 3997* (2013).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).
42. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
43. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

44. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary material for

Illumina But With Nanopore

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

by

Alexander Zee, Dori Deng, Matthew Adams, Kayla Schimke,
Russell Corbett-Detig,
Shelbi Russell, Xuan Zhang, Robert J. Schmitz, Christopher Vollmers

Contents:

Supplementary Figure S1
Supplementary Table S1

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

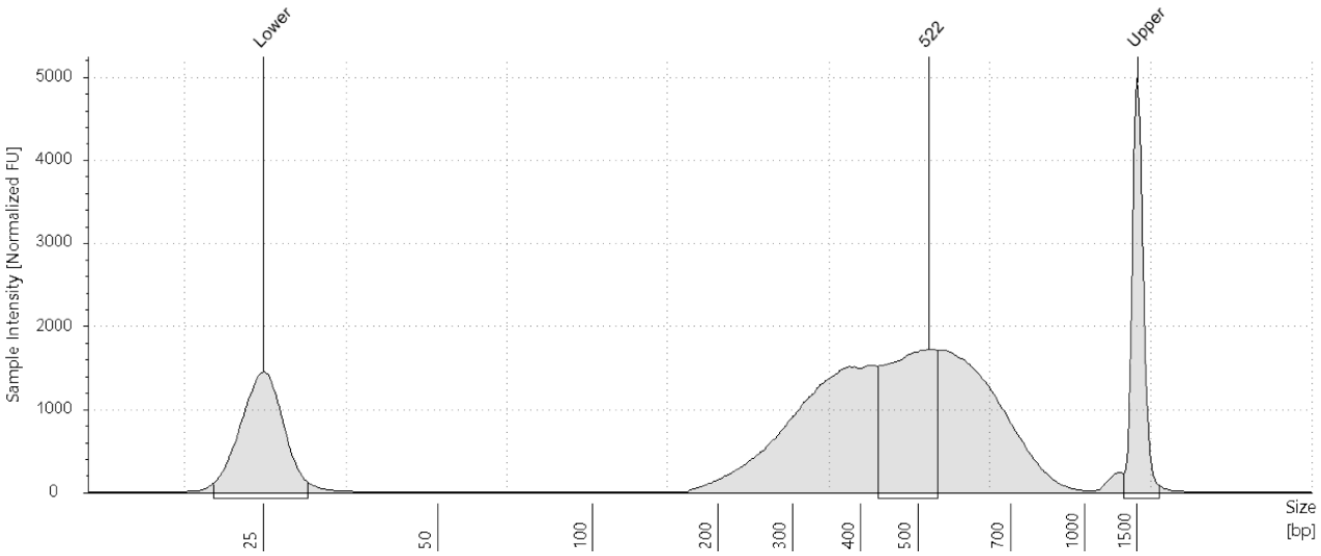

Fig S1. Target-Enriched Tn5 library size.
The size of the target-enriched Tn5 library pool as determined by Agilent Tapestation run.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.30.466545; this version posted October 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Splint oligos

```
>UMI_Splint_1_F_Next_A
GATCTCGGTGGTCGCCGTATCATTTGAGGCTGATGAGTTCCATANNNNNTATATNNNNNATCACTACTTAGTTTTTTGATAGCTTCAAGCCAGAGTTGTCTTTTTCTC
TTTGCTGGCAGTAAAAG
>UMI_Splint_1_R_Next_B
ATCTCGTATGCCGTCTTCTGCTTGAAAGGGATATTTTCGATCGCNNNNNATATANNNNNTTAGTGCATTTGATCCTTTTACTCCTCCTAAAGAACAACCTGACCCAGC
AAAAGGTACACAATACTTTTACTGCCAGCAAAGAG
>UMI_Splint_2_F_Next_A
GATCTCGGTGGTCGCCGTATCATTTGCCGGTTGGGTATCAATAANNNNNTATATNNNNNATTGCCTTTATTCTATCTACTTAGTTTTGGCGATGTAGTCTACCTATCC
TGATGCTGAATAAAGGC
>UMI_Splint_2_R_Next_B
ATCTCGTATGCCGTCTTCTGCTTGAATTAGGTTCTAGGATCACGNNNNNATATANNNNNCTGCCATCGAAAATTTTTCACCCGTAACAAGAACTTACAACTCTCTGAC
GCCTATATCATGAAGGCCTTTATTCAGCATCAGGA
```

Tn5 oligos

```
Tn5ME-R 5'-[phos]CTGTCTCTTATACACATCT-3'
Tn5ME-A (Illumina FC-121-1030): TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Tn5ME-B (Illumina FC-121-1031): GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
```

```
Nextera_Primer_A1   AATGATACGGCGACCACCGAGATCTACAC [i5 index] TCGTCGGCAGCGTCAGATG
Nextera_Primer_B1   CAAGCAGAAGACGGCATACGAGAT [i7 index] GTCTCGTGGGCTCGGAGATGTGTAT
```

Custom blocking oligos for target-enriched Tn5 library prep

```
>NextA_F_Blocking
AATGATACGGCGACCACCGAGATCTACAC  IIIIIIII  TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG/3ddC/
>NextA_RC_Blocking
CTGTCTCTTATACACATCTGACGCTGCCGACGA  IIIIIIII  GTGTAGATCTCGGTGGTCGCCGTATCATT
>NextB_F_Blocking
CAAGCAGAAGACGGCATACGAGAT  IIIIIIII  GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG/3ddC/
>NextB_RC_Blocking
CTGTCTCTTATACACATCTCCGAGCCCACGAGAC  IIIIIIII  ATCTCGTATGCCGTCTTCTGCTTG
```

Table S1: Custom oligos used in the IBWN study