UNIVERSITY OF CALIFORNIA

SANTA CRUZ

LONG-READ SEQUENCING FOR IMPROVING GENOMES AND TRANSCRIPTOMES

A dissertation submitted in partial satisfaction of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

Matthew S. Adams

June 2023

The dissertation of Matthew S. Adams is approved:

_____
Professor Christopher Vollmers, Chair

_____
Professor Angela Brooks

_____
Professor Melissa Jurica

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

LONG-READ SEQUENCING FOR IMPROVING GENOMES AND TRANSCRIPTOMES

By Matthew S. Adams

All the processes of life are controlled by the complex and carefully regulated usage of the genome. Thus, the understanding of an organism's genomic DNA sequence and regions of the genome that are transcribed into complementary RNA transcripts is critical to the study of life and biomedical research. Our understanding of the DNA and RNA composition of a cell has been heavily based on the high throughput short-read sequencing by synthesis technology. However, DNA and RNA molecules, polymers consisting of long chains of molecular subunits, stretch on for lengths far beyond these methods capabilities such that it requires their fragmentation prior to sequencing followed by computational assembly of the short fragments to estimate their arrangement in the much larger original molecule. Thus, I have developed and optimized methods utilizing nanopore based long-read sequencing to improve the accuracy and completeness of genome assemblies and genomic annotations (transcriptome). These methods include a hybrid genome sequencing and assembly workflow that works with minimal amounts of DNA to generate chromosome-scale assemblies, the conversion of short-read libraries for highly accurate nanopore sequencing that makes DNA sequencing more accessible and, an approach for the deep sequencing of full length transcript isoforms to improve genomic annotations for

model organisms. Together, this work improves our ability to understand how living things operate on the molecular level.

Acknowledgements

I would like to thank all the members of the Vollmers Lab for their help and collaborations and creating a wonderful work environment. Additionally, I would like to thank all my other collaborators and co-authors from my time in graduate school. I would also like to thank my family for their support of my education.

Lastly, I would like to thank Chris Vollmers for his dedication to teaching and always putting his students first. He was the best mentor I could have hoped for during my PhD and I will miss being a member of his lab.

Introduction

A Brief History of Sequencing Technology

Determining the precise sequence of residues in nucleic acids is critical in the study of biology and biomedical research. And, while the discovery of the structure and function of DNA was discovered in 1953, the ability to read, in order, the sequence of individual subunits of nucleic acids proved to be a greater challenge. Prior to the 1970's the researchers used a combination of selective nucleases to partially fragment molecules then use biochemical methods to determine the nucleotide composition of the smaller fragments (Heather and Chain 2016). These techniques were only effective for short molecules such tRNA for which the Nobel prize was awarded for the sequencing of the alanine transfer RNA in 1968 (Barciszewska, Perrigue, and Barciszewski 2016). But, the first major breakthrough in sequencing came in 1977 with development of the chain termination method which synthesizes a complementary strand to the target molecule utilizing radiolabelled ddNTPs that terminate synthesis when added to the new strand, fragments are then separated by gel electrophoresis to determine the nucleotide at each position. This method, developed by Fredick Sanger and awarded the nobel prize in 1980, became the dominant method for almost 30 years and is still used regularly today for certain applications. But, even with advanced automation of the Sanger method, the throughput was greatly inefficient for the needs of researchers given the enormous diversity of life.

The second major advancement in sequencing technology came in the 1990's and early 2000's. These methods still relied on sequencing by synthesis like the Sanger method but could be massively parallelized since they did not require gel separation. Instead, these methods detected light emissions from the addition of each base by using a charged couple device (CCD), which are commonly used as sensors in digital cameras. Instruments using this method, developed by 454 Life Sciences and Solexa (later acquired by Illumina), could achieve what would have taken years using the Sanger method in just days or weeks. The increase in sequencing capabilities now grew at a faster rate than advances in computing power as described by Moore's law, that the complexity of microchips (measured by density of transistors) doubles every two years, while sequencing throughput from 2004 to 2010 doubled every five months (Stein 2010). This radically changed biological and biomedical science and ushered in the "genomic revolution" (Shendure et al. 2017).

Third Generation Long-Read Sequencing

The next generation sequencers produced by Illumina have gone on to become the "gold standard" and have obtained a near monopoly within the field. And while they can produce billions of highly accurate sequencing reads at low cost, the technology can only sequence DNA fragments about 150 bp long. These short fragments can be computationally assembled back into the larger original molecule but repetitive genomic elements and alternative transcript isoforms are nearly impossible to resolve with 100% certainty (Steijger et al. 2013).

Third generation sequencing technologies like Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) can produce much longer reads, for PacBio about 10-20 kb, and for ONT reads of 100s of kb and more are possible. These technologies enable repetitive genomic regions to be accurately resolved and entire full length transcripts to be sequenced end-to-end (Byrne, Cole, et al. 2019). Nanopore based sequencing appears to be the most promising third generation technology due to its theoretically unlimited read lengths, ability to directly sequence single molecules of DNA or RNA, direct detection of chemically modified bases, and low instrument costs.

Nanopore sequencing works by using an electrical signal to pass a single molecule through a nanopore embedded in a membrane, as the molecule passes through, the signal is disrupted in a distinct way depending on the molecules sequence which can then be interpreted into the 'A', 'C', 'G', 'T' and 'U' subunits of nucleic acids. But, when ONT introduced the first commercially available nanopore sequencer, the MinION in 2014, the per base read accuracy was less than impressive. Initially, per base accuracy of about 85% was reported, making it difficult to align reads, demultiplex mixed samples with short barcodes, and detect single nucleotide variation. Due to these limitations, the rolling circle amplification to concatemeric consensus method (R2C2), was developed in the Vollmers lab at UCSC (Volden et al. 2018), greatly improved the accuracy of ONT sequencing by sequencing multiple copies of the same molecule then generating a more accurate consensus sequence.

As part of my graduate work documented here, I showed that R2C2 can be used in a hybrid sequencing workflow to cheaply, and easily, generate chromosome-scale genome assemblies using minimal amounts of input DNA. I also show how this method can be used to replace large and expensive Illumina instruments by producing nearly equivalent data on the much more accessible ONT MinION device. And lastly, I generated an isoform level transcriptomic reference for a dozen mouse tissues using exclusively full-length cDNA sequencing.

Chapter 1

One fly - one genome : Chromosome-scale genome assembly of a single outbred *Drosophila melanogaster*

Matthew Adams[1], Jakob McBroome[2], Nicholas Maurer[2], Evan Pepper-Tunick[2], Nedda F. Saremi[2], Richard E. Green[2,3,4], Christopher Vollmers[2,3,5], Russell B. Corbett-Detig[2,3,5]

[This chapter is adapted from a publication **One fly - one genome : chromosome scale assembly of a single outbred *Drosophila melanogaster*** (Adams et al., 2020, Nucleic Acids Research)]

1) Department of Molecular, Cellular, and Developmental Biology, University of California Santa Cruz, CA 95064

2) Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064

3) UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, 95064.

4) Dovetail Genomics, Scotts Valley, CA, 95066

5) Corresponding authors: rucorbet@ucsc.edu, vollmers@ucsc.edu

Abstract

A high quality genome assembly is a vital first step for the study of an organism. Recent advances in technology have made the creation of high quality chromosome scale assemblies feasible and low cost. However, the amount of input DNA needed for an assembly project can be a limiting factor for small organisms or precious samples. Here we demonstrate the feasibility of creating a chromosome scale assembly using a hybrid method for a low input sample, a single outbred *Drosophila melanogaster*. Our approach combines an Illumina shotgun library, Oxford nanopore long reads, and chromosome conformation capture for long range scaffolding. This single fly genome assembly has a N50 of 26 Mb, a length that encompasses entire chromosome arms, contains 95% of expected single copy orthologs, and a nearly complete assembly of this individual's *Wolbachia* endosymbiont. The methods described here enable the accurate and complete assembly of genomes from small, field collected organisms as well as precious clinical samples.

Introduction

The creation of high quality genome assemblies is a key step for the study of organisms on both the level of individuals and populations (Dudchenko et al. 2017). Conventional genome sequencing projects rely on whole-genome shotgun sequencing approaches that generate huge numbers of short sequence reads at low cost. While short reads can be reassembled into larger contiguous genome segments by identifying overlapping reads, they often fail to generate chromosome length assemblies due to the challenge of assembling repetitive DNA sequences. Consequently, many published genomes are highly fragmented (Worley, Richards, and Rogers 2017). Fragmented genomes can be valuable for gene-level studies but many genomic analyses such as understanding chromosome-scale evolution, resolving full-length haplotypes, association studies, and quantitative trait locus mapping require high-quality chromosome-scale assemblies. New hybrid genome assembly approaches can produce highly contiguous assemblies that represent true chromosome length genomes (Rice and Green 2019).

Two recent advances in genomic technologies have dramatically raised the quality of genome assemblies (Yuan et al. 2017). First, third generation long-read sequencing technologies are capable of sequencing entire long repetitive sequences, but they suffer from higher error rates and lower throughput (Worley, Richards, and Rogers 2017). Second, proximity-ligation sequencing, or Hi-C, produces short-read pairs representing sequences that are close together in three-dimensional space (Lieberman-Aiden et al. 2009). This allows high throughput “scaffolding” of

challenging genomic regions (Putnam et al. 2016). However, these impressive gains in genome assembly quality have not been realized across all species due to important biological constraints.

Genome projects can be complicated by the small size of many organisms, which yield corresponding low amounts of DNA from a single individual. Consequently it is not always feasible to obtain sufficient input material for the genomic approaches described above without pooling individuals (F. Li et al. 2019). Nonetheless, developing applications for single individual genome assemblies offers several key advantages. First, it may not be possible to obtain more than a single individual for some species. Second, even if many could be found, pooling several individuals increases the genetic diversity in the DNA input, imposing challenges for accurate genome assembly. For wild caught samples, the possibility of combining cryptic species has the potential to impact assembly quality and introduce spurious biological conclusions. Finally, low input sequencing methods could be used to assemble genomes from precious clinical samples. There is therefore a clear need for new methods that can assemble highly contiguous genomes from a single isolate with limited available DNA.

Recently, Kingan et al. released a whole-genome assembly obtained from a single mosquito, *Anopheles coluzzii,* sequenced using three PacBio SMRT Cells (Kingan, Heaton, et al. 2019). Although the assembly has high contiguity (contig N50 3.5 Mb), the authors were unable to obtain chromosome-scale contigs or scaffolds and the resulting assembly does not include biologically important regions of the

genome that contain chromosomal inversion breakpoints (Kingan, Heaton, et al. 2019; R. B. Corbett-Detig et al. 2019). Additionally, the input material used, approximately 100 ng of high quality DNA, may still be challenging to obtain from a single field-collected individual in many species. Nonetheless, this pioneering work suggests a powerful solution in developing low-input protocols for simultaneously obtaining Hi-C and long-read data from single individuals.

Here, we present a chromosome scale hybrid genome assembly of a single *Drosophila melanogaster* female. From this single individual, we produce long reads, short reads and proximity ligation sequencing data. Our assembly approach leverages the unique value added by each data type to produce a chromosome-scale and accurate genome assembly. This approach is applicable for millions of small species and for irreplaceable clinical samples

Results

Sample Selection

Although numerous studies have assembled genomes from completely (M. D. Adams 2000) or partially (Kingan, Heaton, et al. 2019) inbred arthropods, the genomes of a field collected samples will likely be highly heterozygous outbred individuals. To make our assembly task conservatively challenging yet straightforward to evaluate, we generated an outbred fly by crossing females of the *D. melanogaster* reference strain *y; cn, bw; sp,* or ISO1 (M. D. Adams 2000), to males of another inbred and genetically distinct strain, I38 (Grenier et al. 2015). Importantly, I38's genome is collinear with the reference on broad scales, although smaller rearrangements, such as

small-scale indels and copy number variants, are almost certainly present in the genome (Grenier et al. 2015; Lack et al. 2015). We can therefore use progeny from this cross to demonstrate the applicability of our method for assembling genomes of outbred field-collected arthropod individuals and we can easily verify the accuracy of the assembly by comparison to the ISO1 reference genome. To facilitate the use of several sequencing methods, the single outbred fly chosen for sequencing (referred to as H3) was first laterally dissected (Figure 1).

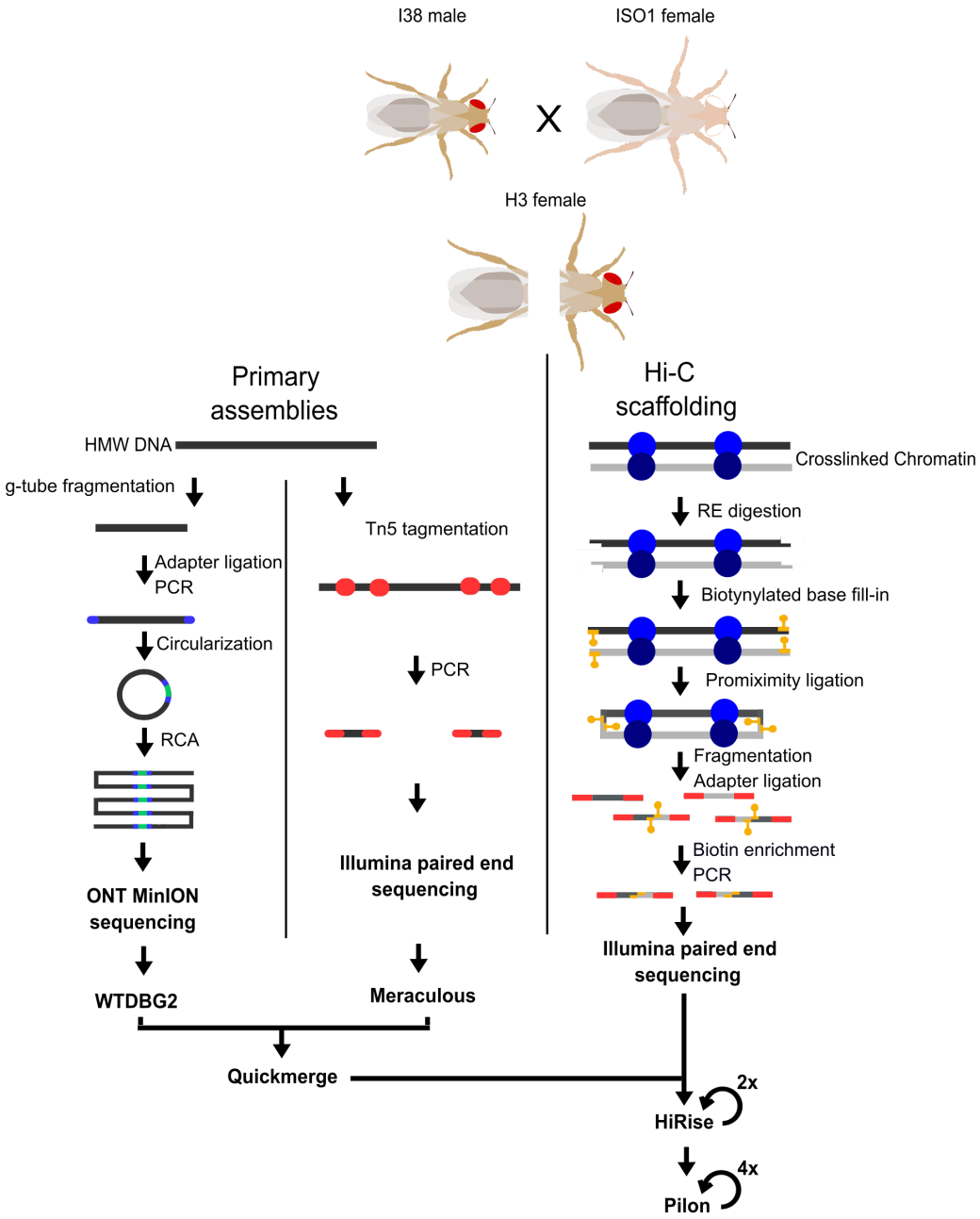


Figure 1. Experimental flow chart. A heterozygous fly (H3) was produced by crossing ISO1 and I38 strains. A single female offspring was laterally dissected. From the posterior half, HMW DNA was extracted and used to prepare the two primary assemblies, a R2C2 genomic library for nanopore sequencing, and a Tn5 tagmentation library for paired end Illumina sequencing. The anterior portion was used to isolate intact chromatin to generate a Hi-C paired end Illumina library. The two primary assemblies were merged into one then arranged into chromosome length scaffolds using the Hi-C contact frequency data.

Primary Sequencing Datasets

From a single outbred adult female fly, we produced short-read shotgun, long-read shotgun and Hi-C libraries (Figure 1). From the posterior half, we extracted high molecular weight (HMW) DNA and we obtained approximately 104 ng in total. We used 78 ng to produce an Oxford Nanopore Technology (ONT) sequencing library following the R2C2 protocol (Volden et al. 2018) with slight modification for genomic DNA (see Methods). The R2C2 protocol generated ONT raw reads that contain tandem repeats of *Drosophila* genomic DNA sequence separated by splint sequences. The R2C2 post-processing pipeline (C3POa) processes these raw reads and generates two types of output reads: 1.) Consensus reads are generated if an ONT raw read is long enough to cover an insert sequence more than once which is evaluated by detecting a splint sequence in the raw read and 2.) Regular "1D" reads for which no splint could be detected in the raw read. In total, 277,305 consensus reads and 1,769,380 "1D" reads were generated from a single ONT MinION flow cell. Both read types were included in the assembly. We additionally produced an Illumina sequencing library using a standard Tn5-based protocol (Methods) and from this we obtained 133,135,777 total paired-end reads (Table 1).

Because both R2C2 and our Tn5 protocol are optimized for low DNA inputs, they require some amplification to produce suitable quantities of libraries for high throughput sequencing. Likely as a consequence, the variance in sequencing depth exceeds the theoretically expected variance if reads were sampled uniformly at random from the genome. Indeed, for libraries with mean depths 236x and 39.7x we

obtained depth variances of 8382 and 1038 for Tn5 and ONT respectively. Nonetheless, we show below that moderately long contigs can still be generated from these data (Supplementary Figure S1).

We also produced a Hi-C library to enable long-range scaffolding across the genome. We optimized a chromatin conformation capture sequencing method (Belton et al. 2012; Lieberman-Aiden et al. 2009) for application to samples with minimal input materials (See Methods). Using this approach and just the anterior half of the fly, we were able to produce 68,400,787 reads in total from a Hi-C library (Table 1). This represents an average of approximately 93,991 clone coverage across the genome. Furthermore, despite low-input, the PCR duplication rate is quite modest (12%). These data therefore indicate that our single-fly Hi-C approach can produce high complexity libraries suitable for scaffolding high quality genomes.

Library	Total Number of Reads	Read Length	Predicted Coverage
Illumina Tn5	133,135,777	151 bp (paired end)	333x
ONT R2C2	2,046,685	3,541 bp (median length)	60x
Illumina HiC	68,400,787	151 bp (paired end)	171x

Table 1. Summary of sequencing data used for assembly and scaffolding

Primary assemblies

To accommodate the unique features of each input data type we produced two primary assemblies. First, we assembled the short-read shotgun dataset using the heterozygosity aware *de Bruijn* graph-based algorithm Meraculous (Chapman et al. 2011). As we are interested in assembling a single haploid genome sequence, we

collapsed the program's resulting diplotigs into a single haploid assembly (*i.e.* "diploid mode 1"). Second, we assembled the processed ONT reads using wtdbg2 (Ruan and Li 2019) (Table 2). As expected given the substantially larger input read lengths, we obtained a much larger contig N50 using this program, than in our short-read based primary assembly (Table 2).

Merging Primary Assemblies

To combine the short and long-read primary assemblies we used the meta-assembler quickmerge. Quickmerge combines two input assemblies to produce an assembly with higher contiguity. Since the input assemblies come from the same individual, gaps in one assembly can be bridged by the other using the alignment of contigs from each input (Chakraborty et al. 2016). The resulting merged assembly had a contig N50 of 274.6 kb (Table 2).

	Contig N50 (Kbp)	**Scaffold N50 (Kbp)**	**Assembly Size (Mbp)**
Meraculous	51	N/A	112.1
wtdbg2	97.7	N/A	112.3
Quickmerge	274.6	N/A	111.2
Hi-Rise	N/A	26,182	111.36
Pilon-Polishing	N/A	26,279	112.22
H3 Genome*	N/A	26,279	110.96

Table 2. Summary of primary and scaffold assembly statistics.*Final assembly size of the H3 fly after removal of the endosymbiont *Wolbachia* genome (see section Genomic Bycatch).

Scaffolding

Although the final merged primary assembly is reasonably contiguous, we observed by far the greatest gains in scaffold size after using our Hi-C data. We ran HiRise to

scaffold the merged primary assembly and a single punitive misjoin was removed before rerunning HIRise a second time (see methods) from which we obtained a scaffold N50 of 26 Mb. Our final scaffolded assembly contains all the major chromosome arms in the *D. melanogaster* genome represented as single scaffolds, and correctly joins arms 2L and 2R across their heterochromatin-rich centromeric region (Figure 2). It therefore appears that the ability to produce high quality Hi-C libraries from extremely limited input material is the most essential component of our method for making contiguous genome assemblies for single individuals in small species.

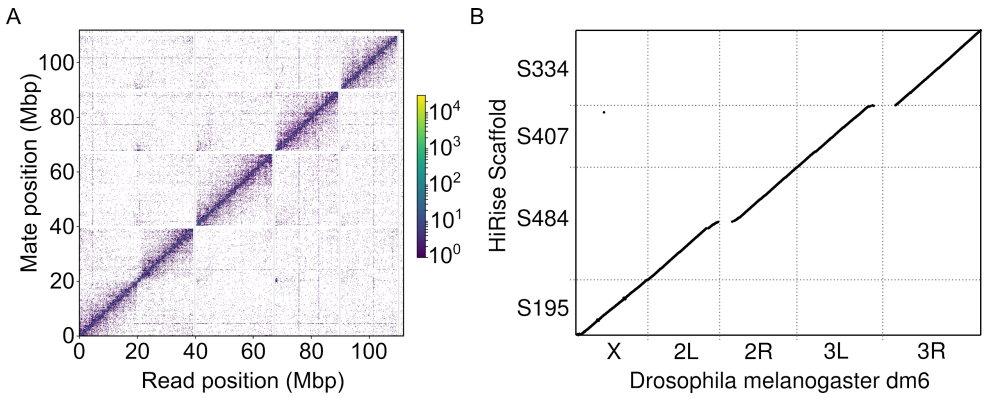


Figure 2: Genome Contiguity. (A) The read density map for Hi-C read pairs mapped onto the five largest contigs in our final assembly. (B) Dot plot of Hi-C scaffold assembly mapped to the dm6 reference genome. Continuous diagonal lines represent full length scaffolds of all major chromosome arms. For clarity of visualization, we restricted this plot to alignments of 5Kb or more using delta-filter in the mummerplot package.

Polishing and Gap Filling

Because we combined diverse data types, and in particular because our primary assembly relies on error-prone long reads, we sought to polish the contigs and fill

gaps in the final highly contiguous assembly. In total we performed four rounds of iterative polishing with Pilon ((Walker et al. 2014), See Methods), until we did not observe significant additional improvements (Supplementary Table S2). The final assembly produced by this step, which we use for all validation below, is the largest of all of our assemblies at 112.2Mb (110.96 Mb after removing *wolbachia* contigs), which presumably reflects the success in our polishing and gap filling by incorporating additional sequences.

Quality of the Final Assembly

We assessed our final assembly quality using several metrics. First, we applied the Benchmarking Universal Single-Copy Orthologs, BUSCO, algorithm (Simão et al. 2015). Briefly, the program provides an assessment of assembly quality specifically with respect to genic sequences by searching for a set of nearly-universal and single copy genes. In applying this quality metric we obtained a BUSCO score of 95.2% completeness for our final assembly. This is slightly lower than the current *D. melanogaster* ISO1 reference BUSCO score of 98.9%, but it is not dramatically different. We therefore conclude that the majority of the expected genic sequences are complete in our assembly.

Second, to compare the assembly of our H3 fly to the dm6 reference and quantify misassemblies we used the genome quality assessment tool QUAST (Gurevich et al. 2013). In addition, we used QUAST to compare another high quality assembly of a different *D. melanogaster* strain, A4 (Chakraborty et al. 2018), to the dm6 reference to set a benchmark for the expected differences between genetically

diverse strains (Table 3). Because A4 was completely inbred and independently isolated from ISO1, whereas our H3 sample is heterozygous for the ISO1 genome, our assembly should more closely match the reference genome. The reason is that we would expect the reference allele to be selected 50% of the time at non-reference sites, and we should therefore observe approximately half as many apparent differences in our final assembly as for A4 relative to the ISO1 reference genome. As expected, our assembly had substantially fewer misassemblies, mismatches and indels than the A4 strain when compared to the dm6 reference, likely because of the relatedness between ISO1 and our assembled individual.

Although our bioinformatic approach has produced a highly contiguous and accurate genome assembly, we acknowledge that alternative approaches might improve on our results. It is typically not possible to extensively optimize a bioinformatic pipeline including all possible variations. We therefore caution that this method should be considered guidelines for processing these types of data, but that researchers should evaluate them carefully for a given assembly task to ensure optimal results can be obtained.

	H3 against dm6 reference	A4 against dm6 reference
# misassemblies	798	2309
# misassembled contigs	15	145
# local misassemblies	1251	3491
# mismatches per 100 kbp	525.36	1136.97
# indels per 100 kbp	88.7	118.84

Table 3. Summary of QUAST output comparing H3 and A4 assemblies to the dm6 reference genome

Repeat Content

Despite similar BUSCO scores and the modest rate of misassemblies that we observe, our genome assembly is approximately 20% smaller than the canonical *D. melanogaster* reference genome. We suspected that much of the difference occurs because our assembly relies on relatively short reads and therefore collapsed repetitive regions. To evaluate this, we used the dm6 annotation data to evaluate coverage across different types of genomic features for both our single-fly assembly and a separate comparison of the A4 assembly. We found that while unique sequence including genes and especially exon sequences were captured in their entirety the majority of the time, highly duplicated elements such as transposons and tRNAs were much less likely to be covered by the H3 assembly (Table 4). This is a general weakness of short-read assemblies (Treangen and Salzberg 2011) and should be acknowledged by any forthcoming analysis applying this method of assembly.

	H3 Assembly	A4 Assembly Control
Coding Sequence (CDS)	94.0%	97.9%
Exon	93.9%	99.5%
Long noncoding RNA	90.6%	98.8%
microRNA	93.7%	99.6%
tRNA	76.5%	98.7%
Mobile genetic elements	55.3%	82.0%

Table 4. Sequence uniqueness strongly impacts assembly coverage. The columns are H3 assembly without any polishing and a non-reference control assembly of standard coverage and size. The rows are annotation types. The value corresponds to the percent of aligned annotated elements with at least 90% of their sequence captured in our assembly. The coverage distribution of our assembly is bimodal, with the vast majority of elements being either covered by a single assembled contig or not covered at all. An expanded table including more annotation types and counts, polished versions of the assembly, and overall assembly statistics can be found in the supplement (Supplementary Table S1).

Phasing

We next evaluated our prospects for phasing the genome of this outbred individual, *i.e.* assigning each heterozygous allele to a chromosome. To do this, we realigned our short-read data to our final genome assembly and called all heterozygous variants using GATK (McKenna et al. 2010). We then realigned the Hi-C and long-read data as well and attempted to infer the phase using combinations of these data and the Hapcut2 algorithm (Edge, Bafna, and Bansal 2017). Because our individual is outbred and we know the complete genome sequence of both ancestors, it is straightforward to quantify the phase accuracy.

Using just the short-read data to phase heterozygous SNPs in the H3 individual, we achieve a modest combined mismatch and switch error rate (*sensu* (Edge, Bafna, and Bansal 2017)) of 0.00147 errors/site. Briefly, mismatch errors

denote sites where single variants are phased incorrectly in an otherwise correct block and switch errors denote a change where at least two subsequent variants are phased incorrectly relative to preceding sites. However the mean phase block length is just 14 heterozygous variants or approximately 2 kb. When we incorporated our Hi-C data, the combined error rate increased to 0.0147 error/site, but nearly entire chromosomes' variants were included in a single phase block (*i.e.*, 99.95% of variants/chromosome). The addition of Hi-C increased switch errors in particular by 0.0126 errors/site. This is likely a consequence of somatic chromosome pairing in dipterans (Cooper 1948), which has previously been demonstrated to create an excess of sister chromosome contacts in Hi-C data (R. B. Corbett-Detig et al. 2019; AlHaj Abed et al. 2019). The increased switch error rate suggests that approximately 17% of Illumina-phasable blocks that are joined by the addition of Hi-C result in switch errors. Therefore, phase inferred from these data could be useful across relatively short distances (*e.g.,* 5 kb), but should be regarded with caution at larger genomic distances. This might not be suitable for all applications of phasing, but would be sufficient for many population genetic questions that rely on short-distance haplotype and linkage information.

Genomic Bycatch

Although not a primary consideration in this work, we found that our assembly captures additional material that is potentially of interest and underscores the power of our approach. First, our selected individual was phenotypically female, nonetheless, we discovered a non-trivial rate of Y-chromosome mapping contigs.

Importantly, we found a similar Y-mapping rate in all three raw sequencing datasets (Supplementary Table S3), and the relevant Y:Autosome depth closely resembles that of typical phenotypic males (unpublished data). We therefore believe this is an XXY female. Despite the abundance of Y-derived reads, our Y chromosome assembly is exceedingly fragmented, as most Y chromosome assemblies are, reflecting the challenges of assembling extremely repeat-dense chromosomes (Kuderna et al. 2019). Nonetheless, this finding highlights the value of sequencing individuals rather than pools because pooling would likely obscure this relationship of relative chromosome depths.

Second, the reference strain is known to harbor the symbiotic bacteria *Wolbachia*, as we used this as the female parent in the cross *Wolbachia* is present in our sample due to infected embryos. Despite the differences in read-depths relative to the nuclear genome, our assembly includes nearly full coverage of the *Wolbachia* genome with few apparent misassemblies (Figure 3 and Supplementary Figure S2). *Wolbachia* in particular (Pietri, DeBruhl, and Sullivan 2016), and endosymbionts more generally (Russell, Chappell, and Sullivan 2019), are frequently present in host somatic tissues, likely explaining the similar abundances of *Wolbachia*-derived reads across sequencing libraries prepared from different parts of the fly. This suggests that in addition to nearly complete nuclear genomes, our assembly method might also be a powerful tool for investigating individual's endosymbiont communities – a fundamental consideration in arthropod biology (Blow and Douglas 2019). Additionally, the analysis of a single individual obviates important concerns about

pooling for interpreting inter-strain endosymbiont diversity (as in, *(Medina, Russell, and Corbett-Detig, n.d.)*), and again emphasizes the potential impact of this approach. See also, Kingan et al for a related approach assembling complete endosymbiont genomes from the genomic data of a single insect (Kingan, Urban, et al. 2019).

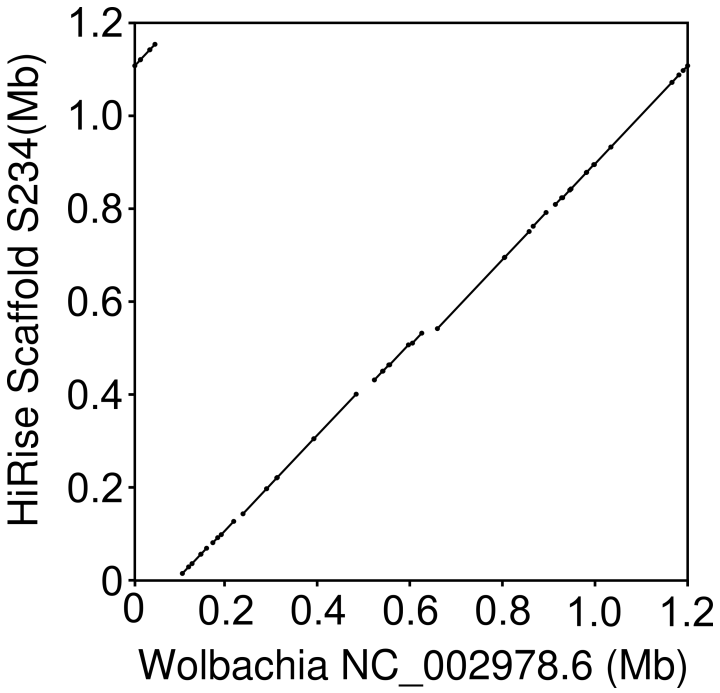


Figure 3. Dot-plot comparison of our nearly-complete *Wolbachia* assembly to the canonical wMel *Wolbachia* genome sequence. Note that the apparent discontinuity in the top right/left, reflects the circular nature of the bacterial genome, and simply indicates that our assembly breaks the circle at a slightly different place.

Conclusion

Recent advances in technology have greatly increased the quality of genome assemblies but generally require a relatively large DNA input. This limitation reduces the applicability of these methods for many precious, rare, and/or field collected

specimens. Here, from a single fly we were able to construct a chromosome scale genome assembly with an N50 of 26 Mb. The primary assemblies were made with less than 90 ng of total input DNA. Therefore, our approach demonstrates that high quality chromosome-scale assemblies can be obtained from limited sample inputs.

Our method also compares favorably for total cost outlay. The DNA isolation and library preparation involves only basic molecular biology methods and equipment. We produced all necessary sequencing data on approximately one half of a HiSeq 4000 lane and a single MinION flow cell. We can therefore produce a contiguous, high quality genome for approximately $1,200 in total materials and reagent costs. For cost effectiveness, our approach compares quite favorably with available alternatives such as Pacbio SMRT cells at $2,000 each.

There are many genome assembly approaches available, and ours may not be optimal for all applications. When input materials are severely limited, the approach we describe here provides an appealing set of trade-offs and may be the only option to produce highly contiguous genome assemblies. Indeed, we have been able to make R2C2 libraries with as little as 10 ng of input DNA. Nonetheless, if more DNA is available, recent advances in PacBio library preparations (Kingan, Heaton, et al. 2019) might be a more appealing option for the long-read assembly. This method does not require amplification, and results in a less biased coverage. However, without Hi-C data for scaffolding, chromosome-scale assemblies are unlikely to be achievable. We therefore consider the addition of our Hi-C approach a necessary prerequisite for high quality genomes.

Perhaps the most fundamental concern for the suitability of our approach is the researcher's specific questions and motivations for making a genome. Applications that require high contiguity in an assembly would be enhanced significantly using this approach. For example, association studies and quantitative trait locus mapping approaches generally require knowledge of large-scale linkage among sites to be successful (Ashton, Ritchie, and Wellenreuther 2017). Similarly, many population genetic frameworks, e.g. those for local ancestry inference (Maples et al. 2013; R. Corbett-Detig and Nielsen 2017), and for estimating past effective population sizes (H. Li and Durbin 2011), are based on the spatial distribution of markers along a reference genome. Finally, comparative studies of large-scale chromosome structure would be significantly enhanced by contiguous genome assemblies (R. B. Corbett-Detig et al. 2019). However, if the distributions of repetitive elements across the genome are of interest, our specific method is unlikely to perform well. Many studies are concerned primarily with coding regions, and for those our approach presents a reasonably high quality option.

This approach can serve as a guide point for genome projects of small organisms which make a large majority of the diversity of life. Approximately 80 percent of known species are insects, and approximately 5 million total insect species are believed to exist on earth (Stork 2018). Additionally, any research projects dealing with minimal DNA could achieve chromosome scale genomic information from this approach. This approach is therefore positioned to revolutionize our understanding of genome structure across diverse species.

Materials and Methods

DNA Extraction

High molecular weight DNA was extracted from one half of a single *Drosophila melanogaster* female using a Qiagen MagAttract HMW DNA kit. One half of a single fly was placed in a 1.5 ml tube with lysis buffer and proteinase k then crushed with a pestle using an up and down motion as to not shear DNA. The lysis and proteinase k digestion was incubated overnight at 37 C. The rest of the purification was performed according to the manufacturer's protocol. The total amount of DNA recovered was 104.4 ng measured with a Thermo Fisher Qubit fluorometer and Qubit dsDNA HS assay kit. This sample was subsequently used for the Tn5 and nanopore library prep.

Illumina Short-Insert Tn5 Sequencing

From the HMW DNA sample, 10 ng of gDNA was tagmented with Tn5 transposase for 8 minutes at 55ºC. The reaction was halted by adding 0.2% SDS and incubated at room temperature for 7 minutes. Four separate PCR reactions were set up using the KAPA Biosystems HiFi Polymerase Kit and amplified for 16 cycles using uniquely indexed i5 and i7 primers. The amplified libraries were pooled and purified using the ≥ 300 bp cutoff on the ZYMO Select-a-Size DNA Clean and Concentrator Kit. 500 ng of the purified library pool was run on a Thermo Fisher 2% E-Gel EX Agarose Gel and cut between 550 and 800 bp. The gel cut was purified with the NEB Monarch DNA Gel Extraction Kit and quantified using the Qubit dsDNA HS Assay Kit and the Agilent TapeStation.

Nanopore Sequencing

From the HMW DNA sample, 78.3 ng was used as input. The sample was first sheared using a Covaris g-TUBE centrifuged for 30 seconds at 8600 RCF. The sheared DNA was size selected using Solid Phase Reversible Immobilization (SPRI) beads at 0.7 beads:1 sample ratio and eluted in 25 ul ultrapure water.

End repair and A-tailing was performed using NEBNext Ultra II End Repair/dA-Tailing Module followed by ligation of Nextera adapters using NEB Blunt/TA Ligase Master Mix following the manufacturer's protocol. The adaptor ligated sample was purified by SPRI beads at a 1:1 ratio and eluted in 50 ul of ultrapure water. The sample was divided into six, 25 ul PCR reactions with Nextera primers and KAPA HiFi Readymix 2x (95 C for 30 s, followed by 12 cycles of 98 C for 10 s, 63 C for 30 s 72 C for 6 min, with a final extension at 72 C for 8 min then hold at 4 C). The PCR reactions were pooled and purified by SPRI beads at a 1:1 ratio and eluted in 60 ul of ultrapure water. Concentration was measured to be 110 ng/ul using the Qubit dsDNA HS assay. The entire sample was size selected by gel electrophoresis using a 1% low melting agarose gel. An area from 6-10 kb was cut out and digested using NEB Beta Agarase I following the manufacturer's protocol then purified using SPRI beads at a 1:1 ratio.

One hundred nanograms of size selected DNA was mixed with 50 ng of a DNA splint and circularized by Gibson assembly using 2x NEBuilder HiFi DNA Assembly Master Mix incubated for 60 min at 50 C. Non circularized DNA was

digested overnight at 37 C using Exonuclease I, Exonuclease III and Lambda Exonuclease (all NEB). Circularized DNA was purified by SPRI beads at a 0.8:1 ratio and eluted in 40 ul of ultrapure water.

The circularized DNA was split into 8 50 ul rolling circle amplification (RCA) reactions (5 ul 10x Phi29 buffer (NEB), 2.5 ul 10 mM dNTPs (NEB), 2.5 ul 10 uM exonuclease resistant random hexamer primers (Thermo), 5 ul DNA, 1 ul Phi29 polymerase (NEB), 34 ul ultrapure water). Reactions were incubated overnight at 30 C. All reactions were pooled and debranched using T7 Endonuclease (NEB) for 2 hours at 37 C. To shear ultra-long RCA products the sample was run through a Zymo Research DNA Clean and Concentrator-5 column and eluted in 40 ul ultrapure water. A final size selection was performed by gel electrophoresis using a 1% low melting agarose gel. An area at approximately 10 kb was cut out and digested using NEB Beta Agarase I following the manufacturer's protocol then purified using SPRI beads at a 1:1 ratio.

The cleaned and size selected RCA product was sequenced using the ONT 1D Genomic DNA by Ligation sample prep kit (SQK-LSK109) and a single MinION flow cell following the manufacturer's protocol. The raw data was basecalled using the Guppy basecaller. Consensus reads were generated by Concatemeric Consensus Caller with Partial Order alignments (C3POa).

HiC Library

The anterior half of the fly was placed into a 1.5 ml tube with 1 ml of cold 1x PBS. 31.25 ul of 32% paraformaldehyde was added. The sample was briefly vortexed and incubated for 30 minutes at room temperature with rotation. After incubation the supernatant was removed and washed twice with 1 ml of cold 1x PBS. 50 ul of lysate wash buffer was added before grinding with pestle. 5 ul of 20% SDS was added then vortexed for 30 seconds and incubated at 37 C for 15 minutes with shaking. 100 ul of SPRI beads were added to bind chromatin. Bound sample was washed 3 times with SPRI wash buffer.

Beads were resuspended in 50 ul of Dpn II digestion mix (42.5 ul water, 5 ul 10x DpnII buffer, 0.5 ul 100 mM DTT, 2 ul DpnII) and digested for 1 hour at 37 C with shaking. Beads were washed twice with SPRI wash buffer and resuspended in 50 ul of end fill-in mix (37 ul water, 5 ul 10X NEB Buffer 2, 4 ul 1 mM biotin-dCTP, 1.5 ul 10 mM dATP dTTP dGTP, 0.5 ul 100 mM DTT, 2 ul Klenow fragment) then incubated for 30 minutes at room temperature while shaking. Beads were washed twice with SPRI wash buffer and resuspended in 200 ul of intra-aggragete mix (171 ul water, 1 ul 100 mM ATP, 20 ul 10x NEB T4 DNA Ligase Buffer, 1 ul 20 mg/ml BSA, 5 ul 10% Triton X-100, 2 ul T4 DNA ligase) then incubated at 16 C overnight while shaking. Beads were placed on a magnet to remove supernatant then resuspended in 50 ul of crosslink reversal buffer (48.5 ul crosslink reversal mix, 1.5 ul proteinase K) then incubated for 15 minutes at 55 C, followed by 45 minutes at 68 C while shaking. Beads were placed on a magnet and the supernatant was transferred

to a clean 1.5 ml tube. 100 ul of SPRI beads were added to the supernatant and allowed to bind before washing twice with 80% ethanol and eluting sample with 50 ul of 1X TE buffer.

The sample was then fragmented by sonication. Fragmented sample was end repaired and adapter ligated using the NEBNext Ultra II kit following the manufacturer's protocol. The sample was purified from ligation reaction by SPRI beads, washed twice with 80% ethanol, and eluted in 30 ul of 1X TE. Biotin tagged fragments were enriched using streptavidin C1 Dynabeads. Enriched fragments were indexed by PCR (23 ul water, 25 ul 2x Kapa mix, 1 ul 10 uM i7 index primer, 1 ul 10 uM i5 index primer) and amplified for 11 cycles. Reaction was purified by SPRI beads and quantified using the Qubit dsDNA HS Assay Kit and the Agilent TapeStation.

Assembly

We produced short-read assemblies using the variation-aware *de Bruijn* graph algorithm, Meraculous (Chapman et al. 2011). Long-read data was assembled using Wtdbg2 (Ruan and Li 2019) using the following options "wtdbg2 -x ont -g 120m -p 0 -k 15 -S 1 -l 512 -L 1024 --edge-min 2 --rescue-low-cov-edges" followed by the wtdbg2 consensus caller wtpoa-cns (Ruan and Li 2019). The two primary long and short-read assemblies were combined using quickmerge default merge_wrapper.py command.

Scaffolding

We polished the hybrid shotgun and long-read assembly using the Illumina shotgun dataset using the bwa mem algorithm (version 0.7.17) (H. Li and Durbin 2009) to map the Illumina reads back to the genome and samtools (version 1.7) to sort the reads. We input the sorted alignment to the consensus for wtdbg (wtpoa-cns) (version 2.5) using the command "-x sam-sr" to polish the contigs of the hybrid assembly. We scaffolded the polished assembly using the scaffolding tool HiRise (version 2.1.1) run in Hi-C mode using the default parameters with the Hi-C library as input. After the first round of scaffolding, we sought to remove putative misjoins in our assembly. To do this, we computed the insulator score across the genome using a 1Mb window on either side of a focal test point. We obtained the expected insulation score for a misjoin between two unlinked contigs by computing the same metric for artificial false-joins between random pairs of unlinked contigs. We then broke the assembly at one aberrantly low insulation score site---indicating little Hi-C support for a specific join consistent with our between contig comparisons.

Polishing

The draft assembly went through a total of four iterative rounds of polishing using the automated software tool Pilon using default settings. For each round the short and long-read data was mapped to the draft assembly using minimap2. After each round, the assembly was evaluated for misassemblies, indels, mismatches, N50,

and assembly size using QUAST (Gurevich et al. 2013) to determine if further polishing would increase the assembly correctness.

Evaluation

To evaluate the completeness of the H3 assembly we searched for conserved genes using Benchmarking Universal Single-Copy Orthologs v3, (BUSCO) with the metazoa odb9 lineage gene set (Simão et al. 2015). To compare to the current reference genome we used the genome quality assessment tool QUAST using the "--large --k-mer-stats" options (Gurevich et al. 2013). Misassemblies are defined by the following criteria, a position in the assembled contigs where 1) the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference, 2) flanking sequences overlap on more than 1 kbp, 3) flanking sequences align to different strands or different chromosomes. Local misassemblies are defined by the following criteria 1) the gap or overlap between left and right flanking sequences is less than 1 kbp, and larger than the maximum indel length (85 bp), 2) The left and right flanking sequences both are on the same strand of the same chromosome of the reference genome.

Repetitive and genic region coverage analysis

We aligned three separate versions of H3 assembly with zero, one, and two rounds of polishing with Pilon to the *Drosophila melanogaster* reference using Minimap2 with default parameters and sam output (H. Li 2018; Walker et al. 2014). We then applied

samtools compression and sorting to produce sorted bam files ((H. Li et al. 2009; Quinlan and Hall 2010)), to which we applied bedtools genomecov with options -ibam and -bga to produce a file of region coordinates and coverage values of 0 or more for each region across the genome (H. Li et al. 2009; Quinlan and Hall 2010) . We combined this information with the annotation gff3 file with a custom script that assigned coverage values to all annotated spans base by base (Quinlan and Hall 2010). The average coverage per base was calculated for each annotated span, then the average and mean value of coverages for all spans for each annotation type was calculated. As a control for comparison we performed this procedure on a complete non-reference *melanogaster* assembly and calculated similar values to elucidate any particular weakness our assembly exhibits.

Phasing

To phase the genome, we realigned all short-read data to our final genome assembly using BWA mem (H. Li and Durbin 2010). We then called all heterozygous variants using GATK (McKenna et al. 2010) on the four largest scaffolds in our assembly, and we filtered this set to exclude SNPs and indels in the bottom 10% or top 10% of observed sequencing depths. As the H3 genome is a mosaic of I38 and dm6 alleles, we "polarized" each heterozygous variant by realigning the dm6 genome using minimap2 (H. Li 2018) to determine whether H3 contained the dm6 allele. We then aligned all Hi-C data using BWA mem (H. Li and Durbin 2010) and the ONT data using minimap2 (H. Li 2018) and attempted to phase the genome using varying combinations of these data using hapcut2 (Edge, Bafna, and Bansal 2017). We

quantified mismatch and switch errors as described in (Edge, Bafna, and Bansal 2017).

Data Access

The sequencing data and final assembly generated in this study has been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA591165.

Acknowledgments

This work was supported by the NIH (R35 GM128932 to RBC-D and R35 GM133569-01 to CV) and from an Alfred P. Sloan Fellowship to RBC-D. During this work JM was supported
by NIH training grant T32 HG008345-01.

Disclosure Declaration

REG is co-founder and paid consultant of Dovetail Genomics.

Supplementary Materials

to

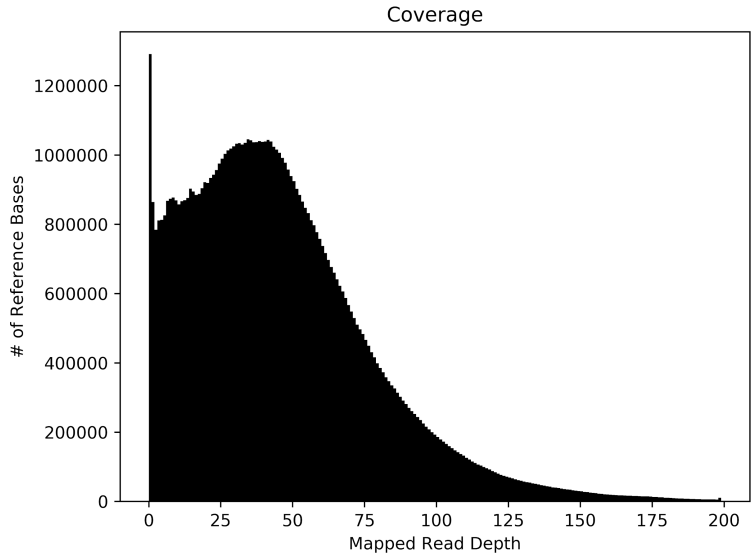
One fly - one genome : Chromosome scale genome assembly of a single *drosophila melanogaster*

by

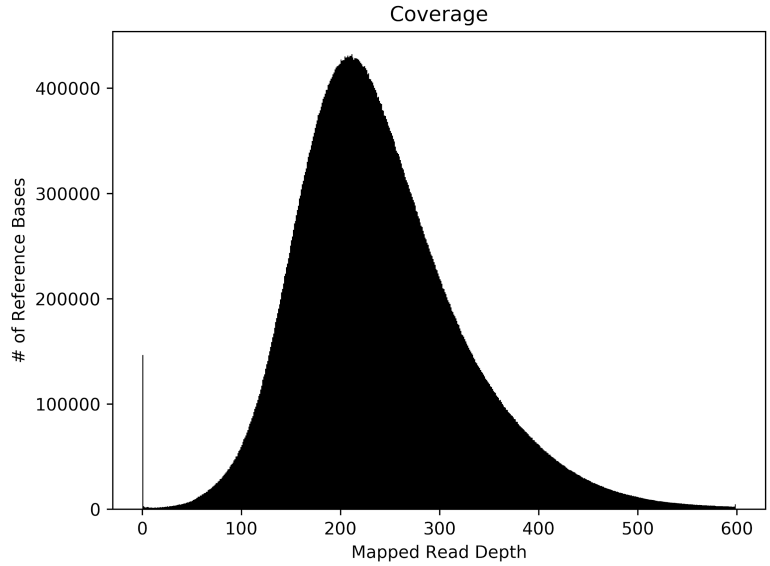
Matthew Adams, Jakob McBroome, Nicholas Maurer, Evan Pepper, Nedda F. Saremi, Richard E. Green, Christopher Vollmers, Russel Corbett-Detig

Table of contents

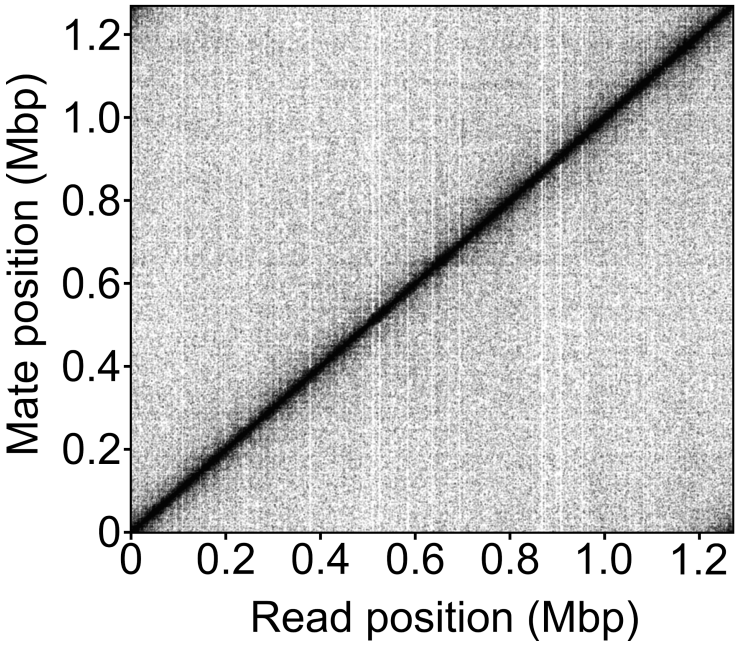
A



B



Supplementary Figure S1: Coverage depth. Sequencing coverage depth of primary assembly data mapped to the dm6 reference genome (A) long-read data set (B) short-read data set. Generated by aligning raw reads to the reference genome using Minimap2 then samtools depth function to calculate depth of sequencing per base of autosomes.



Supplementary Figure S2. Wolbachia Hi-C contact map.

	Pilon0		Pilon1		Pilon2		Pilon3		Pilon4		A4	
Element Type	Mean coverage	Percentage with minimum 90% coverage	Mean coverage	Percentage with minimum 90% coverage	Mean coverage	Percentage with minimum 90% coverage	Mean coverage	Percentage with minimum 90% coverage	Mean coverage	Percentage with minimum 90% coverage	Mean coverage	Percentage with minimum 90% coverage
Mobile genetic element	0.57	52.42	0.59	54.00	0.60	55.02	0.60	55.02	0.61	55.29	0.89	82.03
Gene	0.91	87.92	0.92	88.42	0.92	88.73	0.92	88.73	0.92	88.83	1.10	97.95
lncRNA	0.92	90.06	0.93	90.10	>0.9	90.72	0.93	90.72	0.93	90.68	0.99	98.77
Exon	0.94	93.18	0.94	93.54	0.94	93.81	0.94	93.81	0.94	93.88	1.01	99.54
mRNA	0.93	86.70	0.93	87.43	0.93	87.81	0.93	87.81	0.93	88.05	1.02	97.40
CDS	0.94	93.34	0.94	93.69	0.94	93.93	0.94	93.93	0.94	94.00	1.00	98.97
Primary transcript	0.92	91.63	0.92	92.01	0.93	92.40	0.93	92.40	0.93	92.39	1.66	99.23
miRNA	0.89	92.48	0.89	92.69	0.90	93.52	0.90	93.52	0.90	93.74	1.30	99.58
Pseudogene	0.42	36.69	0.44	37.99	0.44	38.31	0.44	38.31	0.44	38.31	1.08	71.42
tRNA	0.77	74.61	0.80	76.91	0.79	76.48	0.79	76.48	0.79	76.49	0.98	98.75

Supplementary Table S1: Coverage By Element Type. Full table of coverage values for four assemblies for all annotated elements with a total count in the melanogaster genome above 50.

	Assembly size (mb)	N50 (mb)	# misassemblies	# mismatches per 100 kbp	# indels per 100 kbp
Pre-polishing	111.36	26.182	777	671.83	163.30
1st round	111.84	26.24	802	556.68	96.31
2nd round	112.02	26.26	802	532.23	92.40
3rd round	112.14	26.272	800	524.78	89.12
4th round	112.22	26.279	798	525.36	88.77

Supplementary Table S2: Polishing. Brief summary of QUAST genome assembly metrics from four iterative rounds of polishing using Pilon.

	ChrY sequencing coverage	ChrY:Autosome Average sequencing depth
ONT R2C2 library	20 %	2.3x : 46x
Illumina Tn5 library	44 %	40x : 244x
Illumina Hi-C library	37 %	19x : 132x

Supplementary Table S3: Chromosome Y coverage. Coverage of the Y chromosome. Generated by aligning raw reads to the dm6 reference genome using Minimap2 then samtools depth function to determine the percent of Y chromosome coverage and depth of sequencing per base of the Y chromosome.

Chapter 2

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

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*]

[This chapter is adapted from a publication **Sequencing Illumina libraries at high accuracy on the ONT MinION using R2C2** (Zee et al., 2022, Genome Research)]

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

Abstract

High-throughput short-read sequencing has taken on a central role in research and diagnostics. 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 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 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 ONT MinION by using the 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, 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.

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(Mortazavi et al. 2008)), chromatin state (ChIP-seq(Barski et al. 2007) and ATAC-seq(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 things. All of these methods produce a final sequencing library that contains ~200-600bp 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 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. Therefore, placing a benchtop sequencer with capabilities comparable to an Illumina sequencer in most molecular biology and diagnostic labs could be truly transformative by accelerating as well as fully integrating genomics assays into standard lab workflows. In a molecular biology lab, it would speed up developing or establishing new types of sequencing libraries. In a diagnostic lab 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 >20% false negative rate and more accurate sequencing based replacements being available (Ali et al. 2016; Nohr et al. 2019).

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 a successful MinION run can in theory produce 10 million 3kb reads or 5 million 6kb 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 basecalling software guppy5 is only around 96% and 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 R2C2 (Rolling Circle to Concatemeric Consensus) method (Volden and Vollmers 2020; Cole et al. 2020; Byrne, Supple, et al. 2019; Volden et al. 2018; Vollmers et al. 2021a; M. Adams et al. 2020a). 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. 2021a). Since 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, 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, i.e. 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 PLNK (Processing Live Nanopore Experiments), for monitoring and rapid evaluation of sequencing runs. PLNK uses several tools to basecall, 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 percentage 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. We show that for rapid

evaluation purposes, PLNK allows users to observe whether library generation and pooling was successful - enabling Quality Control of libraries often less than an hour into a run. Further, we show that for run monitoring purposes, PLNK makes it possible to evaluate when a predetermined read coverage of a list of target regions is reached. In both cases sequencing runs can be stopped early, saving time and preserving flow cells for future experiments.

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

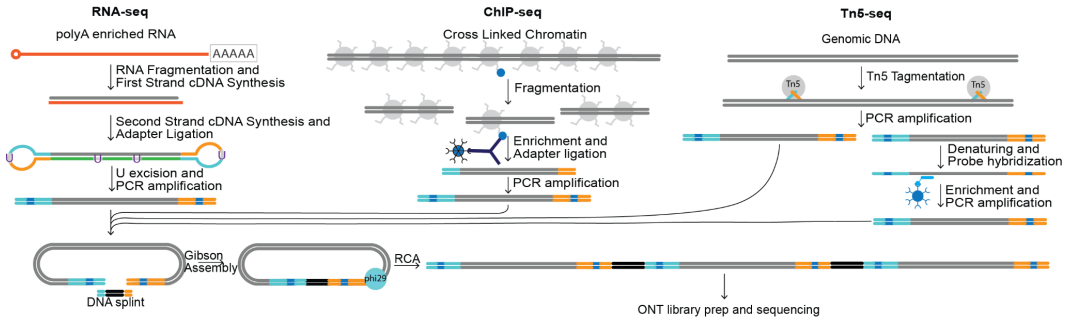


Fig. 1: Experiment overview. Illumina RNA-seq, ChIP-seq, and Tn5-based genomic libraries (regula and enriched) were generated from different samples. The Illumina libraries were then circularized an amplified using rolling circle amplification (RCA). The resulting DNA, containing tandem repeats c Illumina 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 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 (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	Median Per-Read accuracy
RNA-seq	Homo sapiens	9,500,956	2,288	8,992,882	8,066,704	3.14	99.52%
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. For consistency, median per read accuracy is calculated for R2C2 reads prior to demultiplexing.

To benchmark the R2C2 data for the Illumina libraries, we sequenced the same libraries with regular ONT 1D reads and on different Illumina sequencers. We then compared the metrics most relevant to the different library types.

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. We prepared four technical replicate libraries from a single RNA sample in the form of dual indexed paired-end Illumina libraries using the NEBnext Ultra II Directional RNA kit with RNA of the human lung carcinoma

cell line A549. We pooled and sequenced these libraries with the ONT MinION both directly (1D) and after R2C2 conversion (R2C2) as well as with the Illumina MiSeq.

To establish the effect of R2C2 conversion on the throughput of the ONT MinION when sequencing short Illumina libraries, we processed the raw reads generated by both 1D and R2C2 sequencing runs. Raw read numbers for 1D and R2C2 runs generated from one ONT MinION flow cell were similar at ~11.8 million reads. However, 1D reads were less likely than R2C2 reads to 1) pass filter during basecalling, 2) contain both p5 and p7 Illumina adapter sequences, and 3) be successfully demultiplexed. After preprocessing, only 2.5 million 1D reads (21%) remained compared to ~8 million R2C2 reads (Table 2). This means that even a much more productive 1D run, potentially generating up to 20 million raw reads for molecules of this length (F. Pardo-Palacios et al. 2021), would still generate fewer demultiplexed reads (21% of 20 million or <5 million) than the R2C2 run we performed here.

Library type	Raw reads	pass filter	Consensus reads	p5/p7 adapters present	demultiplexed
R2C2 (ONT MinION)	11,789,059	9,500,956	9,132,280	8,992,882	8,066,704
1D (ONT MinION)	11,839,886	7,578,968	N/A	3,469,357	2,530,950

Table 2. R2C2 and 1D read numbers throughout processing steps

To validate the demultiplexing of Illumina library pools from R2C2 data, we compared the ratio of reads assigned to each library in Illumina MiSeq, R2C2, and ONT 1D data based on their combination of i5 and i7 indexes. For all three methods, three technical replicate libraries were pooled at a 4:2:1 ratio. The Illumina MiSeq produced a 4:2.03:1.58 read ratio after demultiplexing. R2C2 produced a 4:1.91:1.34

ratio and ONT 1D produced a 4:2.5:1.82 ratio. With these results being quite similar, the differences are likely due to pipetting variability when pooling the libraries for the different sequencing methods. Further, 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. 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. Finally, 9.71% of R2C2 reads were not assigned to any index combination and 1.7% of R2C2 reads were assigned to index combinations not present in the pool, implying only 0.0289% (1.7%*1.7%) R2C2 reads were assigned to the wrong index combination due to index hopping.

Next we established the effect of R2C2 conversion on read accuracy when compared to ONT 1D and Illumina MiSeq datasets. We aligned all complete p5 and p7 containing and demultiplexed R2C2 (8,066,704) and 1D reads (2,530,950) as well as Illumina MiSeq reads (20,830,560 2x300 bp paired-end reads) generated from these RNA-seq libraries using minimap2. We then calculated the median read accuracy, accuracy per base, and read position dependent accuracy per base (Table 3).

Sequencing method	Median read accuracy (%)	Accuracy per base (%)	Mismatch rate per base (%)	Insertion rate per base (%)	Deletion rate per base (%)
R2C2 (ONT MinION)	99.56	98.87	0.31	0.26	0.55
1D (ONT MinION)	97.2	96.59	1.16	0.81	1.44
Read 1 (Illumina MiSeq)	100	99.47	0.45	0.04	0.04
Read 2 (Illumina MiSeq)	99.54	98.57	1.33	0.05	0.05

Table 3. Sequencing error rates of different methods based on minimap2 alignments of all demultiplex reads

While median read accuracy is a useful and often reported metric to compare error-prone long-read sequencing technologies, it becomes less useful in this study. The sequencing reads we aim to compare are very short - either due to the short length of the molecules sequenced (1D and R2C2)(Fig. 2A) or technology limitations (Illumina MiSeq) - and often accurate enough to be unlikely to contain errors at that length, causing many individual sequencing reads to be 100% accurate. This is obvious with read 1 of the Illumina MiSeq having a median accuracy of 100% which contains little information on the real Illumina MiSeq error rate. Accuracy per base (%), i.e. (correct bases of all reads/all bases of all reads)*100, is a more useful metric to compare accurate short reads. Using this metric we see that 1D reads are the least accurate with an accuracy per base of 96.59%. R2C2 falls between Illumina MiSeq read 1 (99.47%) and read 2 (98.57%) with an accuracy per base of 98.87%. Interestingly, while R2C2 reads contained more deletion and insertion errors, they contained fewer mismatch errors than both Illumina MiSeq read 1 and read 2.

Read position dependent accuracy of 1D, R2C2, and Illumina MiSeq read 1 and read 2 adds further detail to this comparison. In contrast to 1D and R2C2 data, Illumina MiSeq 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). To ensure that our Illumina MiSeq run wasn't an outlier in terms of accuracy typical of Illumina benchtop sequencers, we performed the same position dependent accuracy analysis on publicly available Illumina MiSeq, iSeq, and MiniSeq data, which showed the same overall trends (Fig. S1).

Next, we aimed to establish whether R2C2 RNA-seq and ONT 1D data could be analyzed using computational tools designed and established for Illumina RNA-seq data. To quantify gene expression levels, we aligned and evaluated the entire demultiplexed R2C2 (8,066,704 reads) and ONT 1D (2,530,950 reads) datasets as well as our Illumina MiSeq dataset (20,830,560 read pairs) using the STAR aligner(Dobin et al. 2013) (STARlong executable for R2C2 and ONT1D data) which is routinely used for standard Illumina RNA-seq analysis. 7,365,398 R2C2 reads (91.66%), 1,834,065 ONT 1D reads (72.48%) and 18,649,031 Illumina MiSeq reads (90.08%) mapped uniquely to the human genome, indicating that a larger percentage of ONT 1D reads aren't accurate enough to be aligned by the STAR aligner.

Based on these read alignments, STAR determined normalized gene counts for Illumina MiSeq, R2C2, and ONT 1D datasets. Illumina MiSeq gene counts showed Pearson's r-values of 0.995 and 0.987 when compared to R2C2 (Fig. 2C) and ONT 1D, respectively. Additionally, STAR also determined normalized splice junction counts for the three datasets which provide a higher resolution view of the transcriptome. Illumina MiSeq splice junction counts showed Pearson's r-values of 0.974 and 0.929 when compared to R2C2 (Fig. 2C) and ONT 1D. Finally, we also tested whether ultra-fast pseudo-alignment based tools will generate reliable gene expression levels based on R2C2 and ONT 1D reads which feature more insertion and deletion rates compared to standard Illumina data. We used one such tool, kallisto (17), and found that gene expression values as determined for Illumina MiSeq had

Pearson's r values of 0.985 and 0.973 when compared to R2C2 (Fig. 2C) and ONT 1D.

Overall this comparison showed that using R2C2, we can convert Illumina RNA-seq libraries into DNA ideally suited for the ONT MinION. Not only does R2C2 generate more reads than regular ONT 1D ligation protocols but R2C2 reads are also much more accurate. Because they are more accurate, R2C2 reads are also more efficiently demultiplexed and aligned than ONT 1D reads. Further, because they are similar in accuracy to Illumina reads, standard Illumina tools, like STAR and kallisto, can be used to analyze them. The gene expression and splice junction values generated by R2C2 are highly similar to those generated by Illumina MiSeq data from the same libraries.

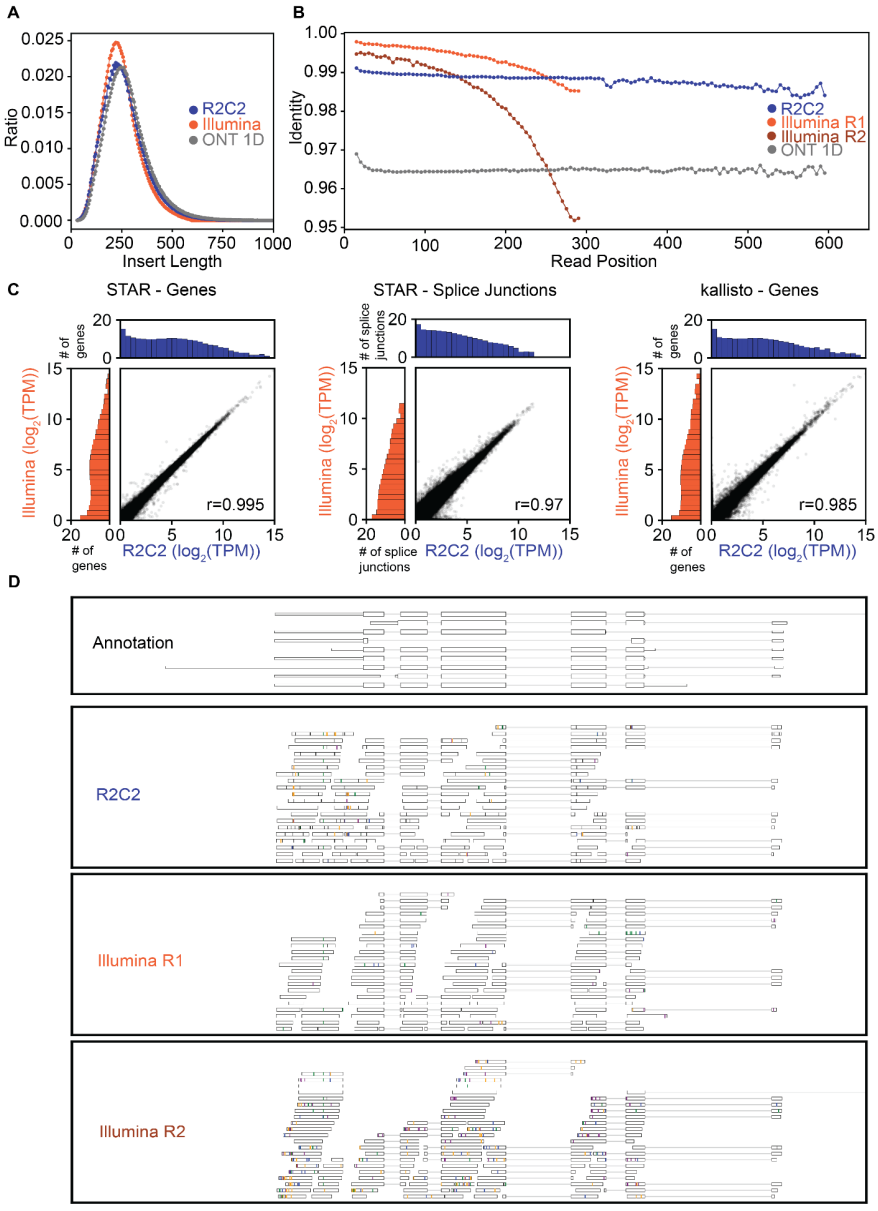


Fig. 2. Sequencing Illumina RNA-seq libraries on the ONT MinION after R2C2 conversion. Insert length distribution (A) and read position dependent identity to the reference genome (B) of R2C2 and Illumina MiSeq reads generated from the same Illumina library. C) Comparisons of R2C2 and Illumina MiSeq read-based 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.

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 (Table 4). Based on their alignment, the sequenced molecule libraries had an insert length of 390 bp (H3K4me3) and 312 bp (Input) (Table 4).

Because the H3K4me3 and Input libraries were 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 the 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 the H3K4me3 and Input libraries (Table 4), respectively, 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 hopping.

The demultiplexed R2C2 reads showed median read accuracy of 99.23% (H3K4me3) and 98.8% (Input) as well as median read length of 556 bp (H3K4me3) and 459 bp (Input) (Table 4). Molecules sequenced by R2C2 were therefore longer than molecules sequenced by the Illumina NovaSeq 6000 (Fig. 3A). The difference between the technologies is likely due to the bias of the Illumina NovaSeq towards shorter molecules.

Sample	Illumina NovaSeq Reads	Median Insert length	R2C2 Reads	Median Insert length
H3K4me3	8,413,865	390	2,493,021	556
Input	32,377,813	312	1,530,914	459

Table 4. 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)(Valliyodan et al. 2019). For alignment, we chose the short-read preset of the minimap2(H. Li 2018) 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(Goodstein et al. 2012) 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 the respective Input reads do so (Fig. 3B).

To investigate this discrepancy in percentage of reads overlapping with H3K4me3 peaks, especially for the H3K4me3 library, we focused on differences between the R2C2 and Illumina sequencing reads. The most obvious difference is the read length with the Illumina reads originating from much shorter molecules (or library inserts). Indeed, when we recalculated this read percentage for Illumina reads originating from inserts longer than 450nt, it increased to 76%. Next, we analyzed the GC content of Illumina and R2C2 reads and found that - in contrast to all other experiments in this manuscript (Fig. S2) - Illumina reads had a lower GC content than R2C2 reads (39% vs 42%). To see whether the difference in insert length and GC content together would explain the discrepancy in percentage of reads overlapping with H3K4me3, we again recalculated this read percentage only for Illumina reads originating from inserts longer than 450nt and with a GC content >39%, i.e. reads derived from long and GC rich molecules. Here, we found that this read percentage increased to 83.2%, virtually matching the R2C2 percentage. Ultimately, this suggested that R2C2 sampled longer and slightly more GC rich molecules from the ChIP-seq libraries. While it is not clear why the longer molecules are more likely to overlap with H3K4me3 peaks, these peaks happen to be more GC rich than the rest of the genome (40% vs 30%) explaining why more GC rich molecules are more likely to overlap with H3K4me3 peaks.

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) (Fig. 3C). This correlation is increased to r=0.866 when this analysis was performed with the longer/more GC rich subsample of Illumina reads but remained lower than what we observed with the RNA-seq data. This means that while R2C2 can be used to evaluate whether a ChIP-seq experiment successfully enriched targeted chromatin, in this particular experiment R2C2 sampled a different population of molecules than the Illumina NovaSeq 6000, thereby complicating quantitative comparisons.

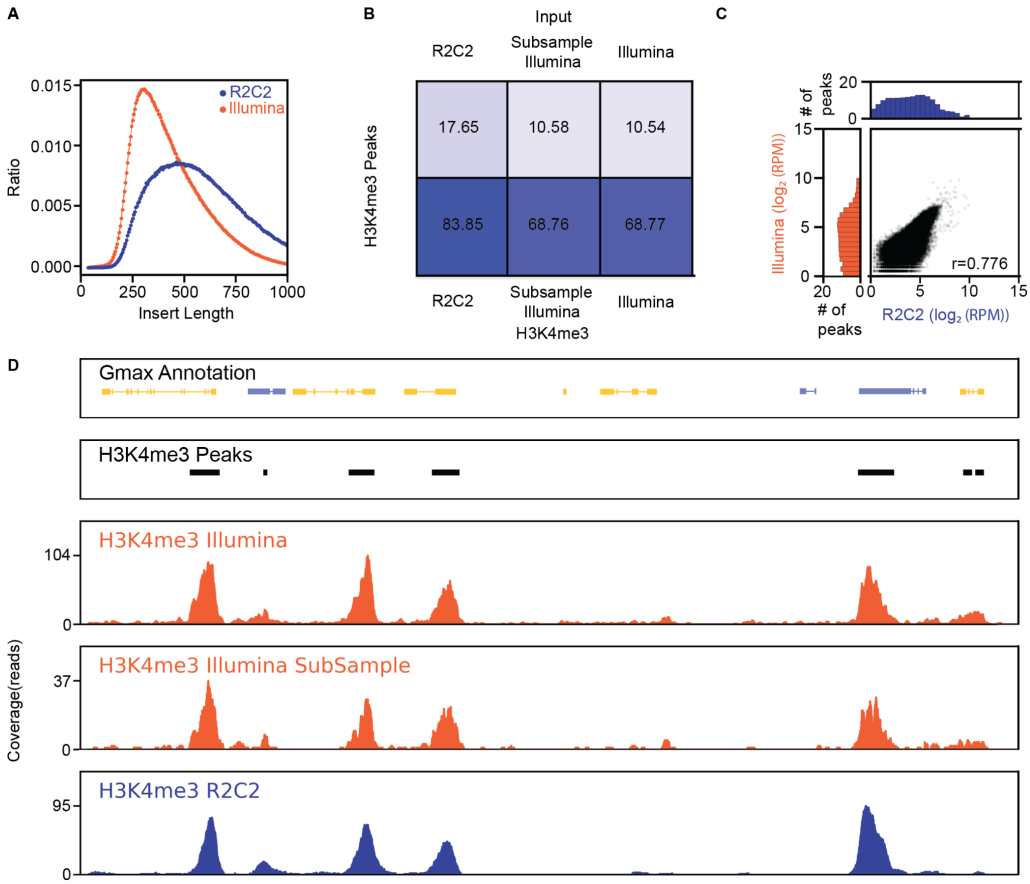


Fig. 3. Sequencing Chip-seq libraries on the ONT MinION after R2C2 conversion. A) Insert length distribution of R2C2 and Illumina NovaSeq 6000 reads generated from the same Illumina library. B) Percentage of reads in the R2C2, Subsampled Illumina and full Illumina datasets overlapping with H3K4me3 peaks 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 section of the Gmax genome.

Evaluating R2C2 for the sequencing of size-selected Illumina Tn5 libraries

In contrast to the other parts of the manuscript which represent head-to-head comparisons between R2C2 and Illumina-based sequencing of the same short-read libraries, here, we tested whether the ability of R2C2 to sequence "medium-length"

molecules >600nt could aid in small genome assembly tasks. Illumina library preparation methods like Tn5-based tagmentation can generate library molecules >600nt which are too long to be sequenced efficiently by Illumina sequencers but can be efficiently processed and sequenced using R2C2. To generate these medium-length molecules for the purpose of genome assembly, we chose to size-select a Tn5-based Illumina 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 library on the ONT MinION.

For this test, we chose to sequence 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. 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(H. Li 2016) 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 (Fig. 4), 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(Chapman et al.

2011). The resulting assembly contained 136 contigs which covered 91.6% of the *Wolbachia* genome (Fig. 4), 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(Gurevich et al. 2013).

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, depending on use case, further polishing with Illumina data might be required to achieve the desired base accuracy.

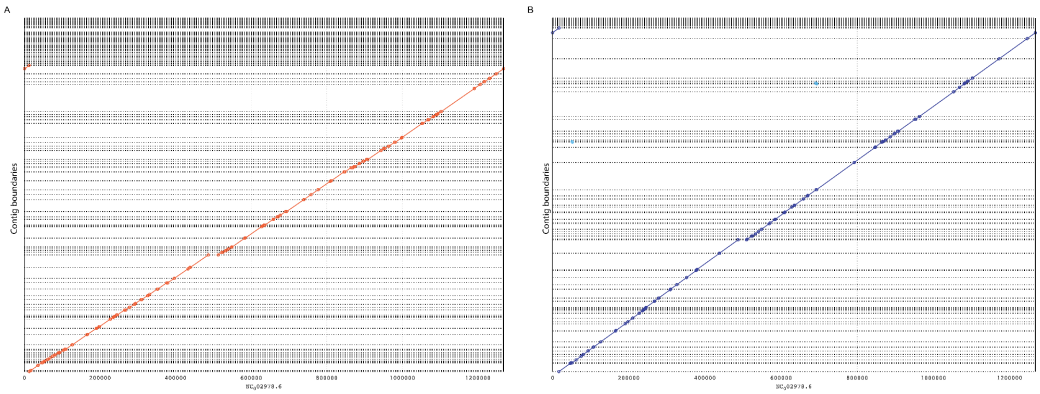


Fig. 4 Comparing R2C2 and Illumina based assemblies of a small genome. Illumina 2x150 reads were assembled in 134 contigs using Meraculous. R2C2 reads were assembled using Miniasm into 95 contigs. The alignments of the contigs of both assemblies - (A) Illumina and (B) R2C2 - are shown as dot plots generated by mummer(Kurtz et al. 2004). Both approaches fail to assemble a section of the Wolbachia genome that contains pseudogenes and a transposable element near to coordinate 500,000.

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(Newman et al. 2014) using a Twist Bioscience oligos panel and reagents (Table S2). We performed this enrichment experiment once, without optimization, and using custom blocking oligos, therefore expecting enrichment to be far from optimal. 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%) R2C2 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 implying that only 0.29% (5.4%*5.4%) of reads were assigned 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(Bushnell, Rood, and Singer 2017). 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 S3). We aligned the reads of different samples and methods to the human genome using the short-read preset of minimap2 and determined the percentage of reads overlapped with a target region and the coverage for each region. For NCI-H1650, 15.8% of R2C2 reads and 14.4% of Illumina MiSeq reads overlapped with a target region producing 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 with 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 fewer 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-H1975 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). Furthermore, the actual accuracy of the MiSeq reads is likely even lower due to alignments not being extended once the 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 pair 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(Poplin et al. 2018) for calling germline variants based on the Illumina MiSeq data and Pepper-DeepVariant(Shafin et al. 2021), a variant caller designed for ONT 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 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 95.9% Precision and 98.3% Recall.

When we visualized the reads on which the False Positive and False Negative R2C2/Pepper-Deepvariant variant calls were made (Fig. S4), we found that the False Positive variants were supported by less than half of the R2C2 reads. Interestingly, when we colored the reads based on the direction of their raw reads, we found that False Positive variants were supported only by reads originating from one raw read direction. We hypothesized that if we oriented reads using the direction of their raw reads - instead of using the p5 and p7 adapters on their ends - before variant calling, it would more closely resemble regular ONT reads and provide more useful information to Pepper-Deepvariant. Indeed, when reanalyzing the reoriented reads and using a QUAL score >=9, Pepper-Deepvariant detected 116 variants which were all present in the Illumina/Deepvariant calls. This means that reorienting the reads before variant calling eliminated all False Positives in the R2C2/Pepper-Deepvariant variant calls. Reflecting known systematic errors of ONT sequencers, two of the three False Negatives missing from the R2C2/Pepper-Deepvariant variant calls were a deletion (TA -> T) next to a 13nt A homopolymer at Chr 17: 7,667,260 and a variant (G->C) next to a 8nt C homopolymer at Chr 12: 120,994,314. The third missing variant, a G->A call at Chr 5: 112,839,666 had a 46% frequency in both Illumina and R2C2 reads, was initially identified as a candidate by Pepper-Deepvariant, but was ultimately scored as a “RefCall”, not a variant. Overall, reorienting the reads by raw

read direction before running Pepper-Deepvariant increased Precision to 100% while achieving a Recall to 97.4%.

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 average per-base accuracy than the Illumina MiSeq. After reorienting R2C2 reads, variants called based on R2C2 and Illumina MiSeq data were very similar. This shows the promise of variant calling based on ONT data but also highlights that extra care has to be taken when preparing data for use in neural network based variant callers like Deepvariant.

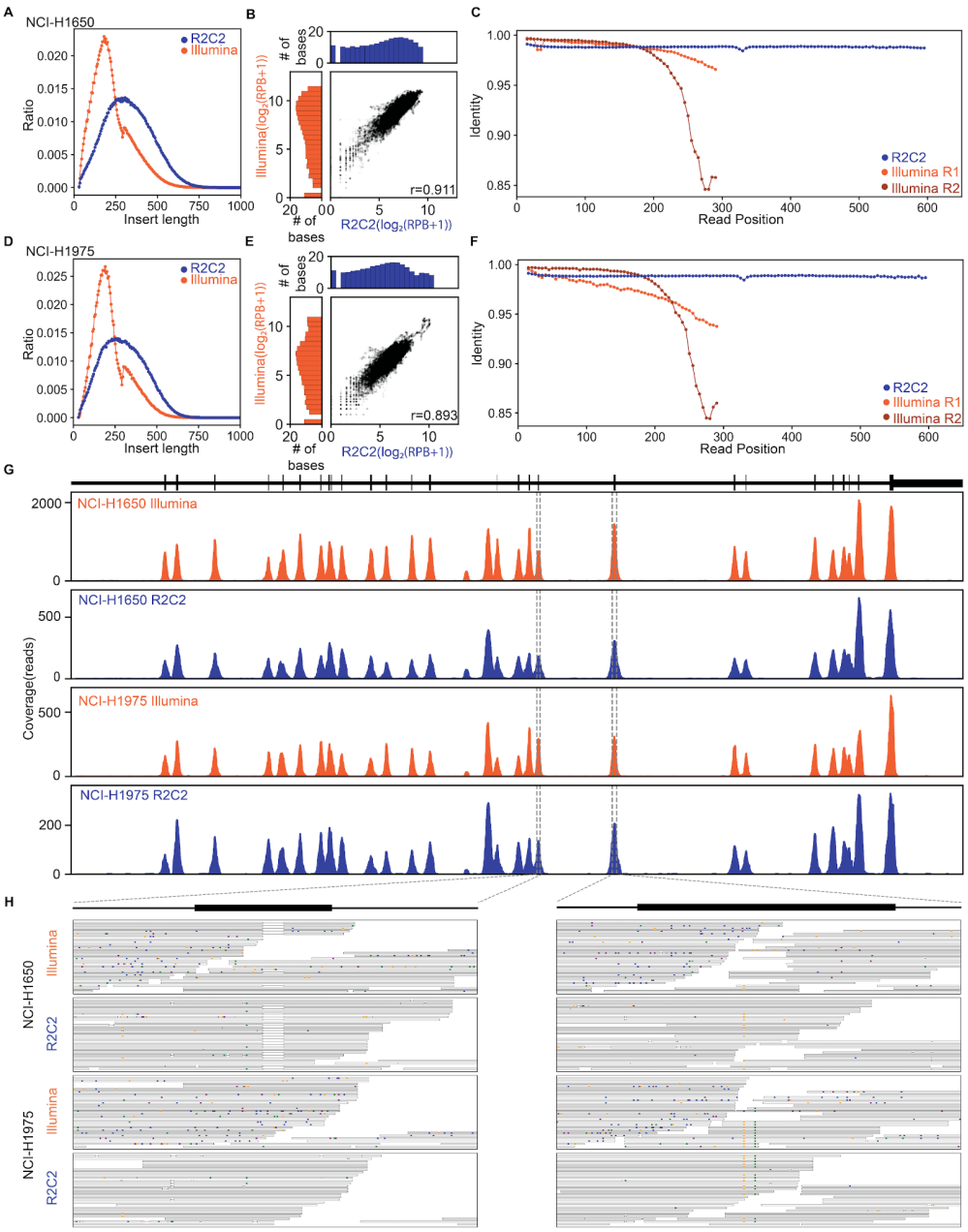


Fig. 5 Evaluating target-enriched Tn5 libraries with R2C2. A and D) Inserts length of library molecules sequenced by Illumina or R2C2 approaches. B and E) Comparison of per-base coverage in

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.
G and H) 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).

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 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 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, 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).

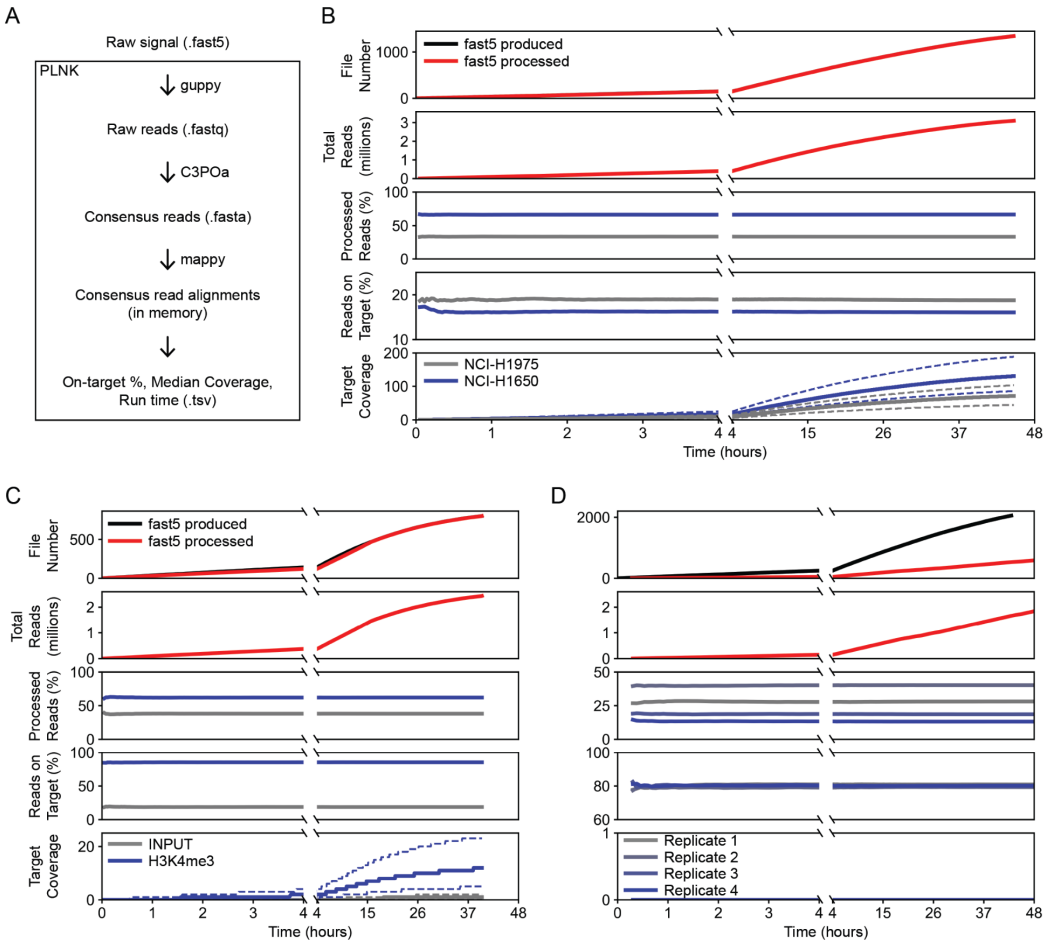


Fig. 6: Real-time characterization of Illumina sequencing libraries. A) Diagram of PLNK functionality, fast5 files processed in the order they are produced. PLNK controls guppy5 for basecalling, C3POa for consensus calling, mappy for alignment, and calculates metrics based on those alignments. B-D) Simulation 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.

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, our Tn5, ChIP-seq and RNA-seq data. 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 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). Additionally, throughout the run, PLNK reported the overall coverage of target regions in real-time.

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.

The bottleneck for analysis in our desktop computer setup seemed to be the guppy5-based basecalling using the slower yet most accurate "sup" basecalling configuration. 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.

Overall, this suggests that PLNK can be used to monitor ONT sequencing runs in real-time. This 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 one 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 flowcell 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.

While long-read sequencers by 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 (Thirunavukarasu et al. 2021; Al'Khafaji et al. 2021) or otherwise preparing (Baslan et al. 2021) 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. 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 <1kb. There is therefore room to optimize ONT library preparations for short read sequencing.

Taking inspiration from the highly accurate but throughput-limited PacBio IsoSeq and HiFi workflows, circularizing-based [R2C2 (Volden et al. 2018), INC-seq (C. Li et al. 2016), HiFRe (Wilson, Eisenstein, and Soh 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. As a result, these R2C2 libraries produced 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 2x300 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 (Table S3). 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 (4-9 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 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. 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 manuscript (Fig. S3), 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 Pepper-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. However, R2C2 requires a library to undergo several processing steps and ONT sequencers feature a unique underlying technology that is totally distinct from Illumina or any other short read sequencing technology. As a result, converting short-read libraries with R2C2 and sequencing them on an ONT sequencer may change what molecules in a pool will be sequenced. 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 where very high read and/or consensus accuracy is required, e.g. 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. Luckily, the barrier-of-entry for R2C2 and ONT sequencing is low so performing a pilot experiment to establish whether R2C2 would be a good replacement for any particular short-read assay should be possible.

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” 600-2000nt 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 3-4 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 (>15kb) 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 RNA-seq, ChIP-seq and enriched Tn5 libraries can be evaluated in under 1 hour 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 lab immediately after a library is produced. The use-cases for this, just as the many use-cases for Illumina benchtop sequencers, will vary from lab to lab. For labs performing small-scale experiments - like RNA-seq of a few samples - the R2C2/ONT MinION combination should be entirely sufficient. For labs performing large scale experiments - like ChIPseq 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 labs could be truly disruptive by eliminating long turnaround times and therefore dramatically 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 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 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

Chromatin immunoprecipitation (ChIP) was performed following the detailed protocol of Ricci et al. with minor modification(Ricci, Levin, and Zhang 2020). 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 was transferred to new tubes. At this point, 10 µl of ChIP input aliquots were collected. Sonicated chromatin was diluted tenfold in the 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 sulfate 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 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 sulfate 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 or directly from Illumina libraries 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). 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.

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 NovaSeq 6000 (ChIP-seq)

Analysis

R2C2 and 1D

Raw nanopore sequencing data in the fast5 file format was basecalled using the "sup" setting of guppy5 to generate fastq files. R2C2 raw reads in fastq format were then processed by C3POa (v.2.2.3 - https://github.com/rvolden/C3POa) to generate accurate consensus reads. R2C2 consensus reads and ONT 1D reads were further processed with C3POa (C3POa_postprocessing.py), using the --trim setting and the following p5/p7 adapter sequences:

>3Prime_adapter

CAAGCAGAAGACGGCATACG

>5Prime_adapter

AATGATACGGCGACCACCGAGATCT

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

RNA-seq

To determine accuracy levels R2C2, 1D, Illumina MiSeq reads were aligned to the human genome reference (hg38) using minimap2 (v2.18-r1015)(H. Li 2018).

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

Position dependent accuracy was determined after converting sam files with the sam2pairwise tool(LaFave and Burgess 2014).

Illumina reads were adapter trimmed using cutadapt (v3.2)(Martin 2011)

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)(Dobin et al. 2013)

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)(H. Li 2018).

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

Peaks in H3K4me3 Illumina data were called using MACS2(Zhang et al. 2008)

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 --secondary=no

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]

We aligned Illumina reads to the Drosophila melanogaster genome (dm6) using bwa mem(H. Li 2013) 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 –secondary=no

Germline variants in Illumina data of NCI-H1650 were called using Deepvariant(Poplin et al. 2018). Germline variant in R2C2 data of NCI-H1650 were called using Pepper-Deepvariant(Shafin et al. 2021)

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.

Analysis of public MiniSeq, iSeq, and MiSeq data

Sequencing runs of genomic E.coli DNA were downloaded from SRA. We selected three runs each for MiniSeq (SRR20643069,SRR20643071,SRR20643072 - generated by the GenomeTrakr project), iSeq (SRR14617007,SRR14617041,SRR14617075) (Mitchell et al. 2022), and MiSeq (SRR19575967,SRR19575968,SRR19575973 - generated by the National Microbiology Laboratory).

To generate accuracy-by-position data, reads for each run were processed separately. First reads were aligned to a E.coli reference genome (CP014314 downloaded from GenBank) using minimap2. Then the genome was then polished using these alignments with pilon (Walker et al. 2014). Reads were then realigned to the polished

genome using minimap2 and position dependent accuracy was calculated after converting the resulting sam files using the sam2pairwise tool.

General Analysis

samtools(H. Li et al. 2009) (v1.11-18-gc17e914) 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 Sequence Read Archive under accession number PRJNA775962

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.

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.

Supplemental material for

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:

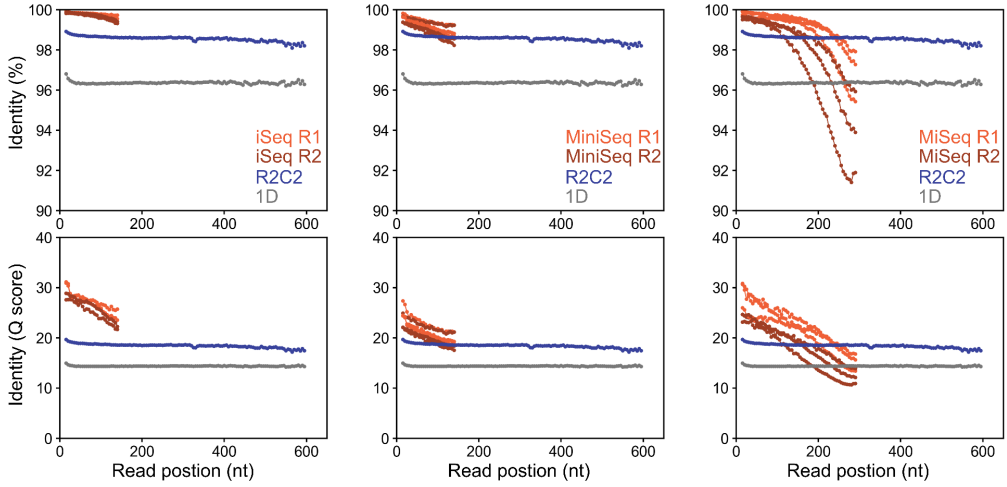


Fig. S1: Read position dependent accuracy of benchtop Illumina sequencers and ONT sequencers. Publicly available iSeq (left), MiniSeq(center), and MiSeq (right) reads of genomic E.coli DNA were processed to evaluate read accuracy. This accuracy, is shown for 3 separate sequencing runs for each read position as percent (top) or log converted Q score (bottom). In each case, Illumina benchtop sequencer accuracy is compared to R2C2 and 1D ONT data generated by us for this study as shown in figure 2B.

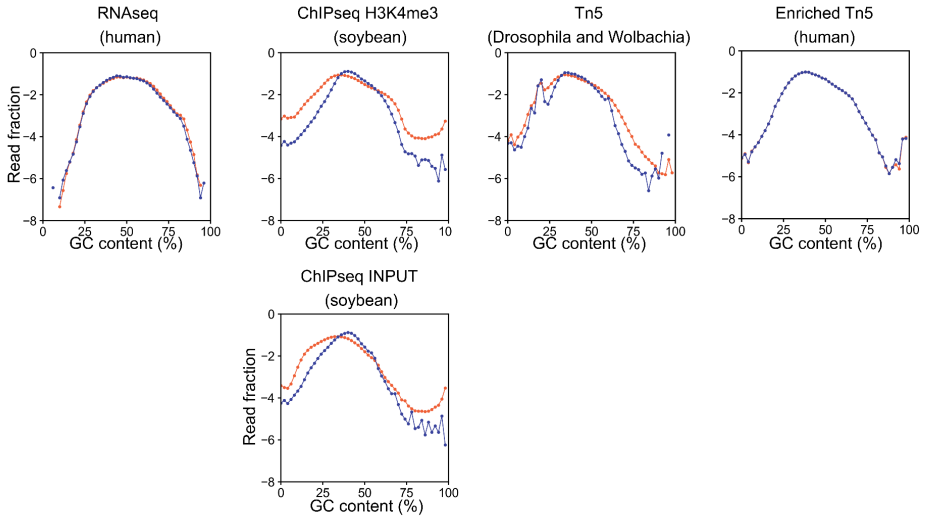


Fig. S2: GC-content of Illumina and R2C2 reads sampling from the same library. Log_{10} converted read fractions of reads with different GC content is shown for all experiments performed for this study. For the ChIPseq study, read fractions for both libraries in the analyzed pool are shown (H3K4me3 and INPUT). Illumina reads are shown in orange and R2C2 reads are shown in blue.

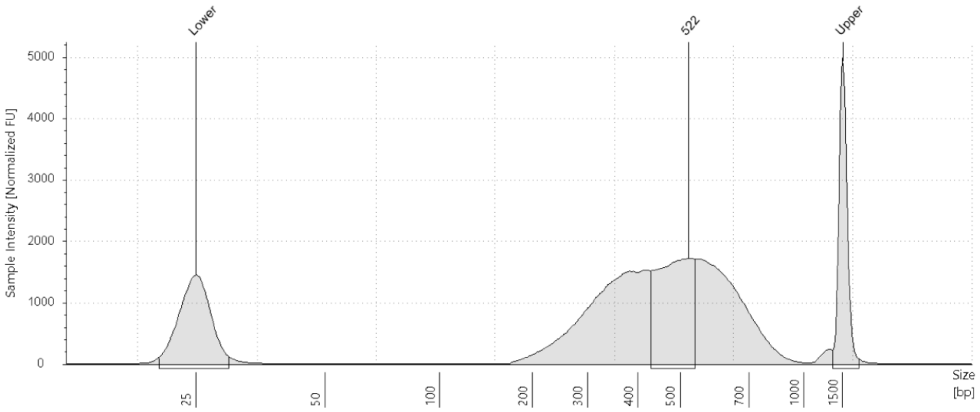


Fig S3. Target-Enriched Tn5 library size.
The size of the target-enriched Tn5 library pool as determined by Agilent Tapestation run.

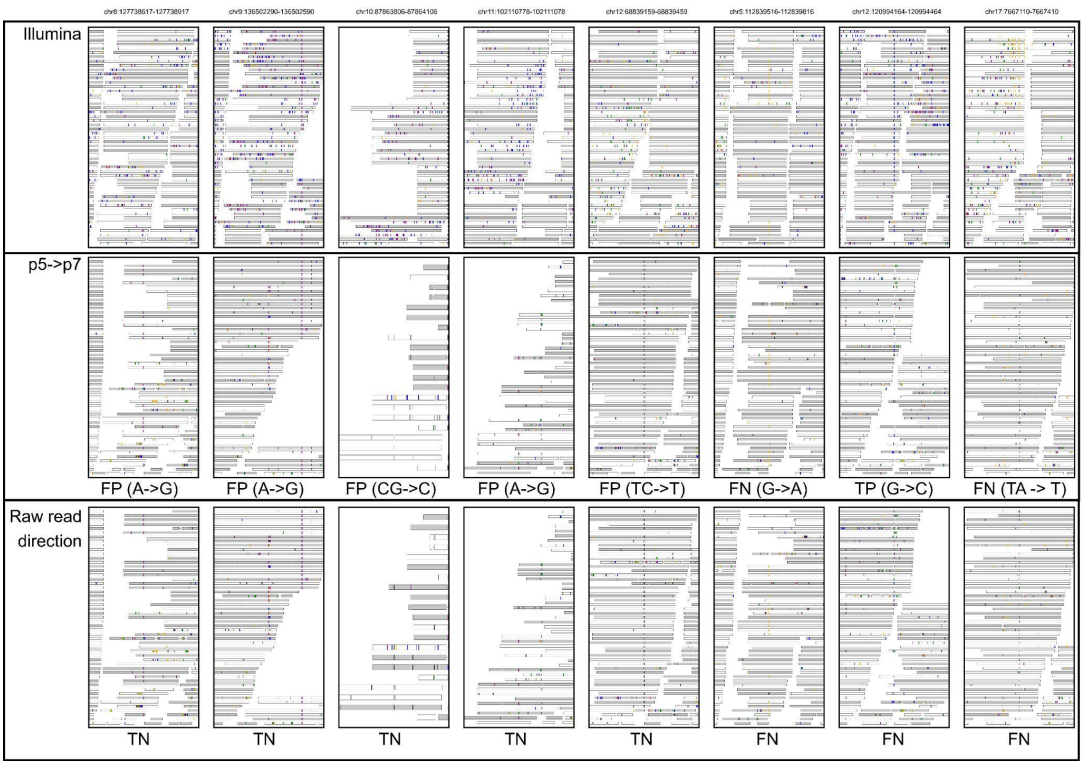


Fig S4: Read context around R2C2/Pepper-Deepvariant miscalls. Subsampled Illumina as well as R2C2 read alignments are shown in genome browser style visualizations around variant calls where R2C2/Pepper-Deepvariant disagreed with Illumina/Deepvariant. R2C2 data and variant calls are shown in both their original orientation (center: p5->p7) as well as reoriented direction (bottom: Raw read direction). Illumina variant call (e.g. A->G) and R2C2 variant call status (FP - False Positive, FN - False Negative, TP - True Positive, TN - True Negative) is indicated for both orientations. In the 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. “Plus” strand alignments are shown with a white background, while “Minus” strand alignments are shown with a grey background.

Splint oligos

>UMI_Splint_1_F_Next_A
GATCTCGGTGGTCGCCGTATCATTTGAGGCTGATGAGTTCCATANNNNNTATATNNNNNATC
ACTACTTAGTTTTTTGATAGCTTCAAGCCAGAGTTGTCTTTTTCTCTTTGCTGGCAGTAAAA
G
>UMI_Splint_1_R_Next_B
ATCTCGTATGCCGTCTTCTGCTTGAAAGGGATATTTTCGATCGCNNNNNATATANNNNNTTA
GTGCATTTGATCCTTTTACTCCTCCTAAAGAACAACCTGACCCAGCAAAAGGTACACAATA
CTTTTACTGCCAGCAAAGAG
>UMI_Splint_2_F_Next_A
GATCTCGGTGGTCGCCGTATCATTTGCCGGTTGGGTATCAATAANNNNNTATATNNNNNATT
GCCTTTATTCTATCTACTTAGTTTTGGCGATGTAGTCTACCTATCCTGATGCTGAATAAAGGC
>UMI_Splint_2_R_Next_B
ATCTCGTATGCCGTCTTCTGCTTGAATTAGGTTCTAGGATCACGNNNNNATATANNNNNCTG
CCATCGAAAATTTTTCACCCGTAACAAGAACTTACAACTCTCTGACGCCTATATCATGAAGG
CCTTTATTCAGCATCAGGA

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

	Read format	Reads/ flowcell	Gbases/ flowcell	Reagents ($)	Reads/ $	Mbases/ $	Machine (k$)
Illumina iSeq 100	150PE	4	1.2	582	6873	2.06	19.9
Illumina MiniSeq	150PE	25	7.5	1750	14286	4.29	65
Illumina MiSeq	300PE	25	15	1750	14286	8.57	99
Illumina NextSeq 550	150PE	400	120	5256	76104	22.83	250
ONT MinION (R2C2)	200-1000SE	4-9	3	650	6154-1385	4	1
ONT PromethION (R2C2)	200-1000SE	20-45	15	1100	18182-36364	13.64	10-75

Table S3: Output and Cost characteristics of R2C2 compared to benchtop Illumina sequencers.

Chapter 3

Mouse Atlas of Tissue Level Transcriptomes

Matthew Adams[1], Christopher Vollmers[2,3,4]

1) Department of Molecular, Cellular, and Developmental Biology, University of California Santa Cruz, CA 95064

2) Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064

3) UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, 95064.

4) Corresponding author: vollmers@ucsc.edu

Abstract

Mice are one of the most commonly used model organisms for biomedical research. Having accurate and complete genomic annotations for model organisms is critical for research efforts. Conventional short-read RNA-seq is unable to accurately capture and quantify gene isoforms due to the initial fragmentation of RNA during library prep. Genome annotation references rely heavily on short-read RNA-seq to create isoform models and thus are incomplete nor do they contain tissue level specificity or quantification of isoform expression. Here, using the nanopore-based R2C2 long-read sequencing method and the Mandalorion isoform tool, we generated a deep, long read, tissue level transcriptome atlas of the BALB/c mouse. This dataset consists of 64 million highly accurate full length cDNA consensus reads, averaging 5.4 million reads per tissue for a dozen tissues. Our mouse atlas represents a valuable reference providing isoform level information for a vital model organism.

Introduction

The mouse has been widely used as a model organism for studying basic biology and biomedical research for almost 100 years. Mice are small, easy to care for, and have short lifespans. Inbreeding of mice has led to genetically identical strains allowing for accurate and reproducible experiments. They share over 15,000 protein coding genes with humans and are susceptible to many of the same diseases (Eppig et al. 2015). Mice are easily genetically engineered to simulate almost any human condition. These features combined make mice critical for scientific research.

For any model organism, a high quality genomic reference and accompanying annotation are invaluable research tools (McGarvey et al. 2015). The initial mouse reference genome was published 20 years ago and recent advances in sequencing technology have improved the completeness and contiguity of genome assemblies to point of true telomere to telomere genome assemblies (Mouse Genome Sequencing Consortium et al. 2002). But complete reference annotations of functional DNA elements are a more complicated challenge. Many projects have set out to work on this task ((Frankish et al. 2019),(Kawai et al. 2001), (McGarvey et al. 2015), (ENCODE Project Consortium 2004)). The most basic level of genome annotation defines on the nucleotide level the location of coding and non coding genes and regulatory elements. For multicellular eukaryotic organisms creating complete genome annotations requires the analysis of every cell type using a number of different experimental methods due to unique gene expression patterns and post

transcriptional processing (Morillon and Gautheret 2019). As a consequence, references are incomplete due to limitations of sequencing technology.

Short read RNA-seq has become the gold standard for understanding transcriptomes due to its high throughput nature, and ability to quantify gene expression and exon inclusion (Mortazavi et al. 2008). This has made it the most common method for genomic annotations. Short read RNA-seq requires the fragmentation of RNA during library prep but even the most advanced computational methods fail to assemble short reads into full length isoforms (Steijger et al. 2013). Consequently this method is unable to provide a comprehensive isoform level view of the transcriptome.

Third generation long-read sequencing technology such as ONT and PacBio are capable of generating sequencing reads of many kilobases and even up to megabase long reads (Byrne, Cole, et al. 2019). For the application of RNA-seq this means that entire full length transcripts can be captured as single reads that include the poly(A) sites, transcription start sites (TSS), and splice sites without the need to assemble short fragments. Genome annotation research using long-read sequencing ((Sharon et al. 2013), (Glinos et al. 2022)) have shown the value of capturing full length transcript isoforms but also the limitations in regards to throughput, accuracy, and molecule length bias during library prep and sequencing.

Here, we have utilized the nanopore based R2C2 long-read sequencing method ((Volden et al. 2018), (M. Adams et al. 2020b),(Vollmers et al. 2021b), (Zee et al. 2022)) and the Mandalorion isoform calling pipeline (Volden et al. 2022) to

create a high quality isoform level transcriptome reference for 12 tissues from the BALB/c mouse strain to determine what isoforms are expressed for each gene, at what relative expression level, and how isoform usage varies across tissues.

Results

Sequencing Overview

A tissue level, long-read transcriptome atlas was constructed using commercially available high quality RNA (Takara) from 12 mouse tissues (brain, eye, heart, kidney, lung, liver, salivary gland, smooth muscle, spinal cord, spleen, testis), each pooled together from dozens to hundreds of male and female balb/c mice. Sequencing libraries were prepared using a modified Smart-Seq2 protocol and oligo(dT) primers (see methods). To increase sequencing coverage of longer transcripts, which are biased against during multiple steps in the library preparation and sequencing process, some of the cDNA was size-selected for molecules 2 kb in length and above by gel electrophoresis.

Non size-selected and size-selected libraries were prepared for sequencing using the R2C2 protocol and sequenced on a combination of Oxford Nanopore Technologies MinION and PromethION sequencers (R9.4 pore chemistry and SQK-LSK110 library preparation kits). After basecalling using Guppy (v.5) we generated accurate full length cDNA consensus reads using the C3POa pipeline. In this way, we produced 64 million full length cDNA consensus reads, averaging 5.4 million reads per tissue. For non size selected libraries the median insert length was approximately 750 bp while the size selected libraries had median insert length approximately 2kb. Together, the

distribution of non size-selected and size-selected read lengths reflected the length of likely full-length transcripts in GENCODE basic vM30 annotation. Further, the full-length R2C2 consensus reads were very accurate, with the median per base identity for non size selected and size selected reads being 99.8% and 98.9%, respectively (Figure 1).

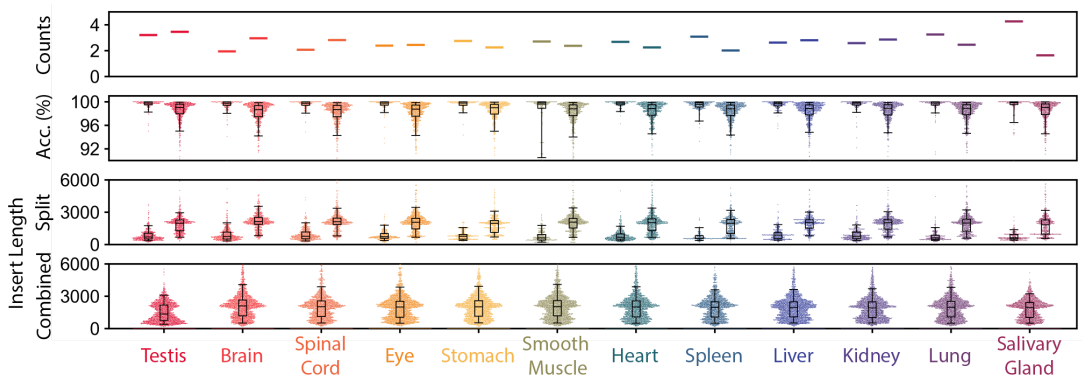


Figure 1. Overview of data set. Top panel, read counts in millions split between non size-selected and size selected libraries. Second from the top, read accuracy of C3POa full length consensus reads split between non size-selected and size-selected libraries. Second from bottom, insert length split between non size-selected and size selected libraries. Bottom panel, insert length of combined non size-selected and size selected libraries.

Gene Level Analysis and Illumina Comparison

Next, we determined whether the full-length cDNA R2C2 reads we generated could be used for gene level analysis. The first analysis we performed aimed to determine if our sequencing depth was enough to capture all annotated genes (GENCODE) present in the sample (Figure 2A). We compared gene saturation of our data set to publicly available Illumina data (right panels) generated from the same RNA for ten of the twelve tissues we sequenced (all but stomach and testis)(PMID: 25730492

). Comparing data sets on the level of individual tissues (top panels), we see the R2C2 data set approaching a plateau but with fewer total genes than the Illumina data which can be at least partially attributed to the significant difference in read counts. The combined data sets for both methods each identify just under 40,000 annotated genes.

A key output of RNA-seq experiments is the quantification of gene expression by counting the number of reads that align to a particular annotated gene. We compared gene expression quantifications produced by the R2C2 method to the more standard short read Illumina sequencing data using the same RNA. We compared tissue level combined datasets of size selected and non size selected reads (Figure 2B). Pearson correlation gave an r value of ~0.85 across tissues (only brain shown) indicating a strong correlation of gene quantification between the two methods.

We also investigated the number of unique annotated genes detected by either R2C2 and Illumina sequencing. Because the illumina data contained significantly higher read counts we randomly subsampled 3 million reads from each tissue for both methods to achieve a more balanced comparison. We grouped subsets of genes identified by each method (Figure 2C). The majority of genes detected were identified by both methods and the method that identified the most genes varied depending on the tissue.

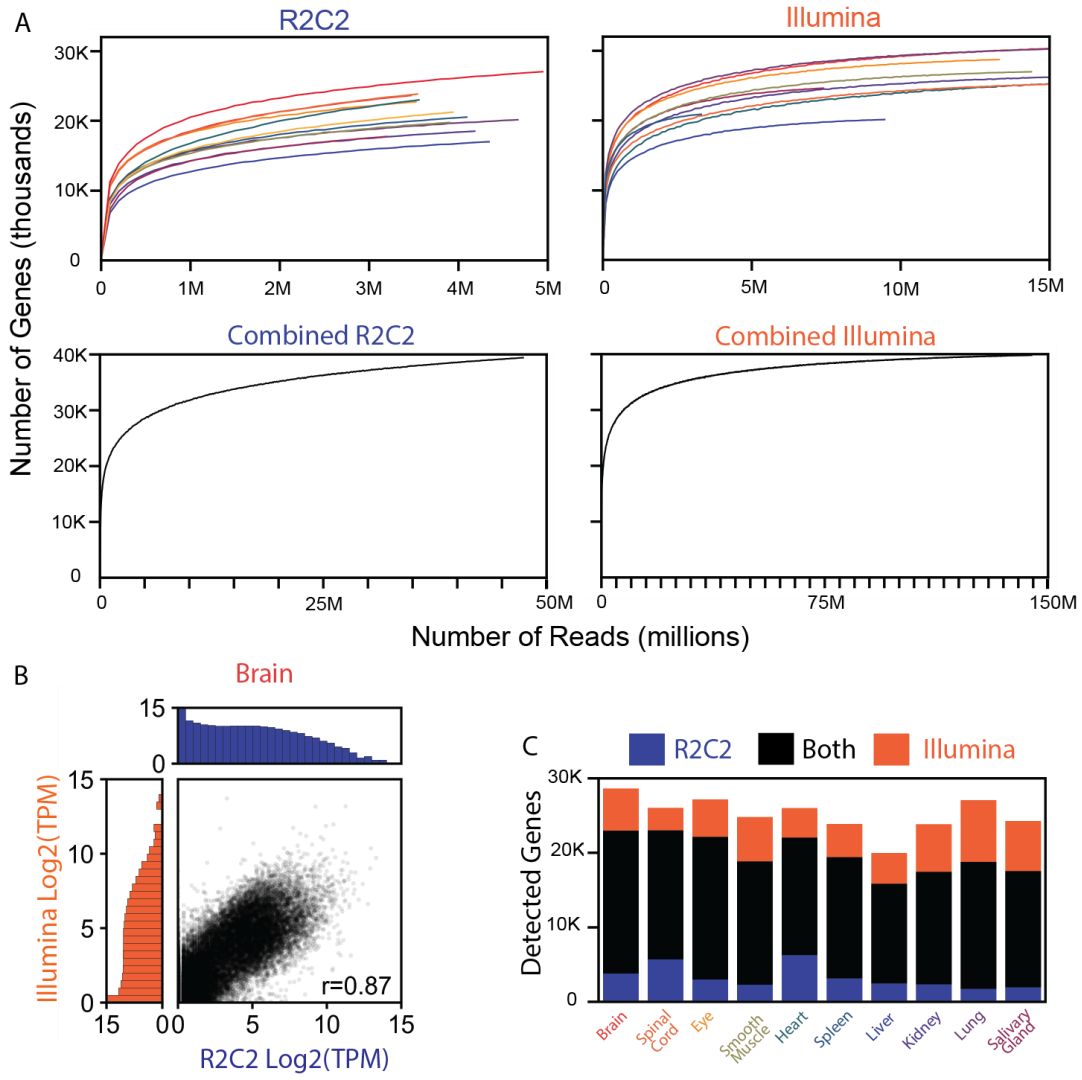


Figure 2. Gene level analysis. Panel A, gene level saturation curve analysis, left side, R2C2 data, right side, Illumina data, top panels, individual tissues, bottom panels, all tissues combined. Panel B, scatter plot of gene counts from R2C2 vs. Illumina for brain. Panel C, Comparison of genes detected by either R2C2 or Illumina.

Isoform Characterization

Next, we aimed to use the full-length R2C2 consensus reads to move beyond gene level analysis and define comprehensive isoform annotations for each of the 12 tissues in this study. To identify isoforms in a way that is highly specific, i.e. reports

few false positive isoforms while retaining high sensitivity for previously annotated and new isoforms, we analyzed the R2C2 reads we produced using Mandalorion (v4.0) tool. Mandalorion takes as input a 1) reference genome (fasta) and 2) annotation (gtf, optional) along with full length transcript sequence data (R2C2 or Iso-Seq). It aligns reads to the genome using minimap2 (H. Li 2018) and groups reads into isoforms and creates a consensus for each isoform. Mandalorion then aligns isoform sequences and filters the consensus isoforms. We ran Mandalorion on both the entire data set, as well as on individual tissue data sets. When run on the entire combined dataset, Mandalorion produced 167,079 isoforms. The median isoform length was approximately 2kb (Figure 3A) and median isoform accuracy 99.97%.

When run on the individual tissue data sets, Mandalorion identified between 22,727(salivary gland) and 63,948 (testis) isoforms. To determine whether we sequenced these data sets to exhaustion, i.e. more reads would not result in more isoforms being identified, we performed a saturation analysis for each tissue which also highlights the transcriptional complexity of each tissue (Figure 3B). Because we did not reach saturation for most tissues, we wanted to determine that we at least identified one isoform for each expressed gene identified in the illumina data. We found that on average across tissues, for ~80% of the genes detected by Illumina that had greater than 1 TPM, we had at least one isoform in our data, and ~91% for genes with greater than 5 TPM (Figure 3C).

This initial isoform-level identification and analysis showed that our consensus isoforms have very few errors and a length distribution that matches

closely with GENCODE annotations. While we did not reach isoform level saturation, we were able to capture at least one isoform for the majority of genes.

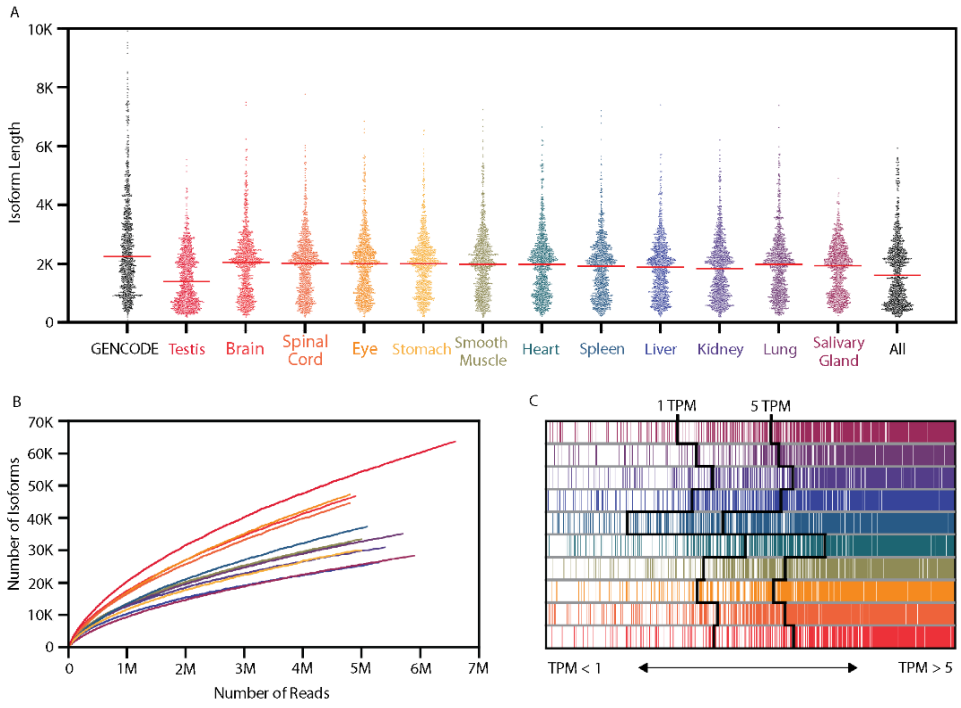


Figure 3. Isoform Characterization Panel A, Isoform length distribution for each tissue and the combined data set compared to GENCODE vM30 basic protein coding transcripts. Panel B, Isoform level saturation curves for each tissue. Panel C, R2C2 isoforms that match genes detected by Illumina, sorted from lowest to highest TPM from Illumina data, left most black bar marks 1 TPM, right most black bar marks 5 TPM.

Isoform classification

Next we aimed to further classify the isoforms we had identified. To this end, we used SQANTI3 (Tardaguila et al. 2018) to characterize the full length consensus isoforms Mandalorion had identified for each tissue. When we analyzed the combined dataset, SQANTI3 detected 68,947 genes, 25,331 annotated genes and 43,616 unannotated genes. Approximately 80% of annotated genes had more than one isoform and ~30%

had over 6 isoforms. For novel genes only ~8% produced had more than one isoform. 86% of isoforms produced by novel genes contained just one exon, compared to ~20% of isoforms produced by annotated genes.. This indicates that many Mandalorion isoforms identified as novel genes by SQANTI3 can likely be attributed to genomic contamination, antisense transcripts from known genes, or transcribed enhancer regions.

SQANTI3 classifies isoforms into four main categories based on comparison to the reference annotation file: full splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC), novel not in catalog (NNC). For FSM, the isoform must have the same number of exons and matching splice junctions, while the exact 5' and 3' ends can differ, ISM have fewer terminal exons but still match annotated splice junctions, NIC uses some combination of known splice junctions and, NNC have at least one splice junction that is novel. Across tissues we see approximately 50% of isoforms called by Mandalorion classified as a FSM to GENCODE annotation and about 80-90% falling into the four main categories at similar ratios with the exception of testis with a disproportionate number NNC isoforms indicating poor annotation of isoforms unique to testis in GENCODE. (Figure 4). When analyzing the data from individual tissues using SQANTI3, we found that the testis expressed the highest number of genes and isoforms followed closely by neural tissue (brain, spinal cord).

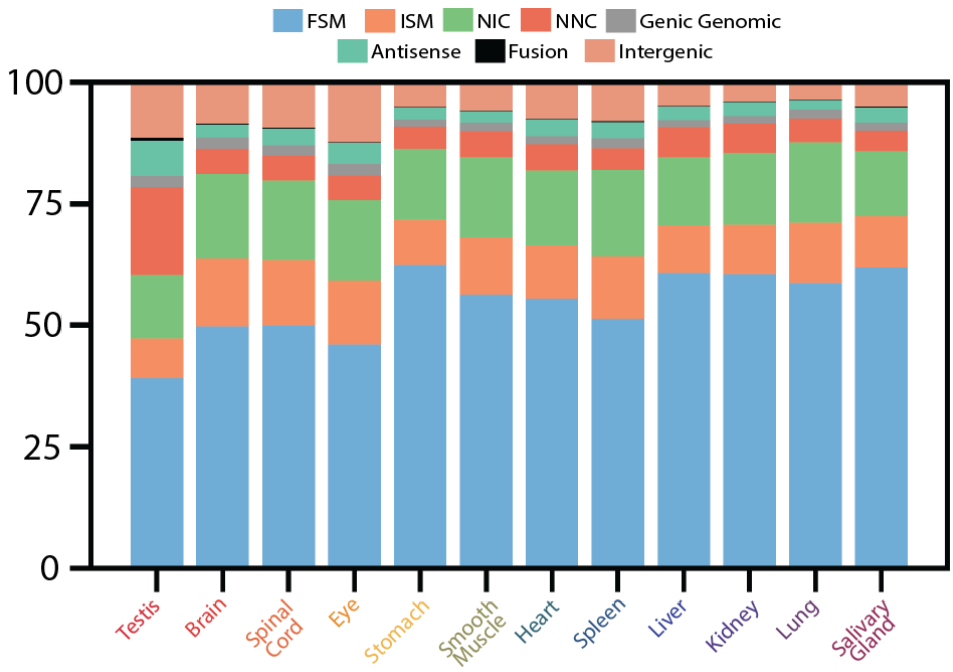


Figure 4. Isoforms Classification. SQANTI3 isoform classification categories for each tissue.

Differential Isoform Usage Across Tissues

Next, the quantitative nature of the R2C2 approach as well as the multiplexed setup of our sequencing strategy allowed us to compare isoform expression across tissues. To do so, we used the isoforms identified from the combined data set for which Mandalorion quantified the expression in each tissue.

To identify genes with differential isoform expression across tissues, we performed a Chi-squared contingency table test. We found 5,654 genes that had significant differential isoform usage (p-value < 0.05) with a minimum of 50 reads in at least two tissues. An example of one such gene, Rab3il1 shown in Figure 5, highlights differential isoform usage across tissues particularly in regards to the use of

alternative TSS and first exons, as well as alternative internal exon usage within the same tissue.

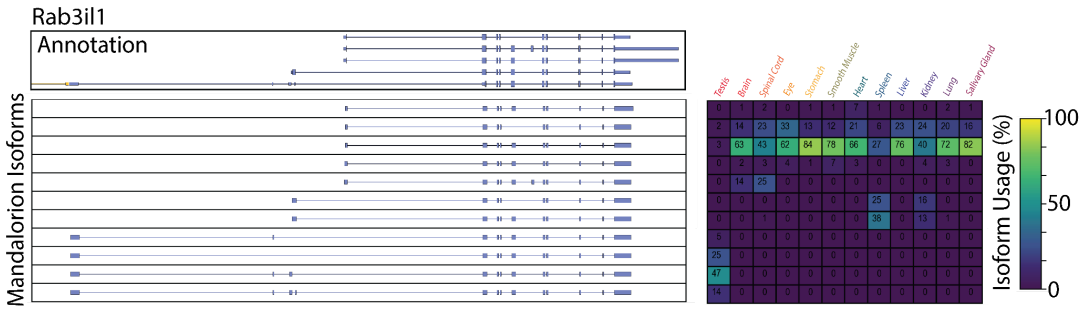


Figure 5. Differential Isoform Usage. Genome Browser shot of Rab3il1 is shown with GENCODE v30 annotation on top and isoforms called by Mandalorion below. Right side, relative usage of each isoform in each tissue, yellow indicates higher usage, blue indicates lower usage.

Novel Genes and Isoforms

We investigated sequences determined to be novel genes or isoforms by Mandalorion. Less than 1% of reads in the whole data set were assigned to a novel gene and almost half of those were assigned to just five novel genes . Further investigation revealed the five most highly expressed novel genes were the results of unannotated gene duplications (Mucl1, Ahsp, and a ribosomal LSU gene) and unannotated pseudogenes (Cox7 and Gm12338).

Testis

Testis are known to be the most transcriptionally complex tissue in mammals in terms of the number of expressed genes and isoforms (Kaessmann 2010) and for this reason we chose to further investigate the nature of isoform expression in testis. In our data, we found 63,948 isoforms expressed in testis and over half, 32,843, had unique

transcription start sites compared to any other tissue (+ or - 100bp), significantly higher than any other tissue which averaged just over 3,000 TSS not used in another tissue. We then investigated if the use of alternative TSS could be validated by transcription factor binding data. To do this, we used available transcription factor ChIP-Seq data from Chip-Atlas.org (Oki et al. 2018) that included data for 4 different TFs [Rfx2, Taf7l, Tbpl1,Mybl1] expressed in the testis. We found that ~33% of the testis unique TSS were within a TF ChIP peak (Figure 5, bottom panel).

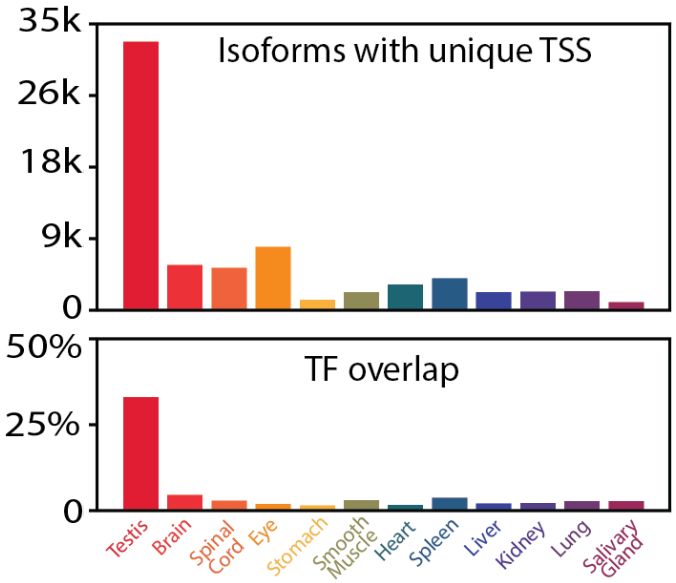


Figure 6. Validation of Unique TSS Top panel, the total number of isoforms with unique TSS for each tissue. Bottom panel, percentage of isoforms with unique TSS from each tissue that had TSS overlap with TF ChIP data from ChIP-atlas.

Data Sharing

Unlike existing genomic reference annotations, our data set includes isoform level expression for genes across tissues. To create an easily accessible resource for

researchers needing isoform level expression data we have created a custom genome browser session using the UCSC genome Browser (Navarro Gonzalez et al. 2021) (https://genome.ucsc.edu/s/vollmers/MouseAtlasOfTissueLevelTranscriptomes). This session contains Mandalorion isoform models and quantification which provides researchers with information on what isoforms are present for their gene(s) of interest and their relative usage.

Discussion

Here, we have presented a high quality isoform level tissue transcriptome atlas for BALB/c mouse. We sequenced approximately 64 million full length transcripts across a dozen tissues averaging 5.4 million reads per tissue and covering almost all high and medium expressed transcripts and a significant portion of lower expressed transcripts. This data set will be a valuable resource for scientists requiring detailed isoform level information for their genes of interest. Researchers can easily look up the highest expressed isoform for a particular gene of interest in a given tissue and have it synthesized for use in functional assays and other experiments. Current annotations are unable to capture this information due to their reliance primarily on short read RNA-seq that is limited in its ability to identify and quantify transcripts at the isoform level. As long-read sequencing technology continues to improve and gain acceptance we will see great improvements in the completeness and accuracy of genome annotations for all organisms. Indeed, GENCODE has developed the TAGENE pipeline to incorporate long-read data sets into their transcript models, resulting in the addition of new genes and transcripts to their most recent annotations

(Frankish et al. 2023). Additionally, the Long Read RNA-seq Genome Annotation Assessment Project (LRGASP) consortium (F. J. Pardo-Palacios et al. 2022), launched by GENCODE and other partners, has systematically evaluated various long read sequencing methods and computational analysis tools (including R2C2/Mandalorion) and will certainly advance the adoption of long-read RNA sequencing.

Methods

Sample Multiplexing

RNA was acquired from Takara (Cat# 636644). Multiplexing samples was done using one of two methods, the first used barcoded DNA splints for Gibson assembly then pooling samples after rolling circle amplification, the second method used barcoded oligo(dT) for cDNA synthesis wich allowed pooling before Gibson assembly. Both methods produce equivalent data. Approximately 80% of the data used in this study was generated by using barcoded oligo(dT) primers for multiplexing tissues. *Library Preparation and Sequencing*

RNA was first mixed with dNTPs and oligo(dT) primer, either barcoded or non-barcoded, then denatured to remove secondary structure for 3 minutes at 72C. First strand reverse transcription (RT) using Smartscribe Reverse Transcriptase (Clontech) and SmartSeq template switching oligo (TSO) with DTT and SUPERaseIN was performed for 1 hour at 42C then heat inactivated for 5 minutes at 70C. Second strand synthesis and PCR with KAPA 2x master mix and ISPCR primer with RNaseA and lambda exonuclease for 12 cycles3 (37C for 30 minutes, 95C for 3

minutes, 98C for 20 seconds, 67C for 15 seconds, 72C for 8 minutes, 72C for 5 minutes, 4C hold). cDNA was cleaned up and size-selected using SPRI beads at a 1:0.85 (sample:beads). After quantification by Qubit the cDNA libraries were pooled together if barcoded oligo(dT) primers were, if not, cDNA from individual tissues would still be kept separate. The cDNA was then split for size selected and non size selected R2C2 library preparation. For size selection, cDNA was run on a 1% low melt agarose gel and everything over 2 kb was excised and purified using beta-Agarase digestion and SPRI bead clean up. Size selected and non size selected cDNA were further processed separately but identically. cDNA libraries were circularized by gibson assembly (NEBbuilder HiFi) with a short DNA split that overlaps with the ends of the cDNA. For cDNA that was not barcoded during cDNA synthesis a barcoded DNA split was used. To remove un-circularized molecules, a linear exonuclease digestion with ExoI, ExoII, and Lambda Exonuclease (all NEB) was carried out for 16 hours at 37C then heat inactivated for 20 minutes at 80C. The reaction was then cleaned using SPRI beads at a 1:0.85. The clean, circularized library is then used as a template for rolling circle amplification (RCA) using Phi29 (NEB) with a random hexamer primer for 18 hours at 30C then heat inactivated for 10 minutes at 65C. The phi29 reaction was then debranched using T7 endonuclease for 2 hours at 37C before being cleaned and concentrated using Zymo DNA clean and concentrator column. The library was quantified by Qubit and gel extracted as described above but the region extracted was a bright band just over the 10 kb marker. After gel extraction, the library was quantified again by Qubit. Libraries barcoded

during the Gibson assembly step were now pooled together at equal mass. We used the Genomic DNA by Ligation (SQK-LS110) kit from ONT to prepare for sequencing following the manufacturer's protocol. The final library was loaded onto either a ONT MinION or PromethION sequencer. Flowcells were nuclease flushed and loaded with additional library partway through sequencing based on pore availability statistics shown in the MinKNOW software to increase sequencing throughput.

Data Processing

All ONT fast5 files were basecalled using Guppy (v.5) with the super accurate configuration. R2C2 full length consensus reads were generated and demultiplexed by C3POa (v2.4.0). Read trimming was performed using a custom python script. Isoforms were called by the Mandalorion Isoform analysis pipeline (v4.0) run on both individual tissue data and the combined data set.

Analysis

Per base identity was determined by aligning reads to the mouse reference genome (GRCm39) and parsing the alignment using a custom script.

Gene level saturation curves were produced by random subsampling of featureCounts output for each tissue and the combined dataset. Isoform level saturation curves were produced by random subsampling fasta files output by C3POa and running Mandalorion on each subsample.

Scatter plots comparing Illumina and R2C2 gene quantification were produced using gene read counts from featureCounts and the figure was generated from a custom

python script. R2C2 data was aligned to the reference genome using minimap2 while Illumina was aligned using STAR aligner.

Mandalorion isoforms produced from both individual tissues and the combined dataset were used as input for the sqanti_qc.py script of SQANTI3 v5.1.

Differential isoform usage analysis was performed using Chi2 contingency test with a custom python script utilizing SciPy.

A tissue unique TSS was determined using a custom python script that compared every isoforms TSS from each tissue to the entire dataset. A tissue unique TSS was defined as a TSS in one tissue that did not exist in any other tissue + or - 100 bp. A custom python script was used to determine TSS overlaps with publicly available ChIP-seq (https://chip-atlas.org/). Data from chip-atlas.org was downloaded as a BED file from the peak browser tool by selecting the following options: Assembly: M. musculus mm10, experiment type: ChIP TF, Cell Type Class: Gonads, Threshold for Significance: 50, ChIP Antigen: all, Cell Type: testis.

Acknowledgements

The ONT Promethion sequencing was carried out by the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01.

Funding

This work was supported by the NIH R35 GM133569-01 to CV

Data Availability

The full length consensus reads generated in this study have been submitted to the NCBI Sequence Read Archive under accession number PRJNA971991

UCSC genome browser tracks can be found at,

https://genome.ucsc.edu/s/vollmers/MouseAtlasOfTissueLevelTranscriptomes

Conflicts of Interest

The authors declare that they have no conflicts of interest with the contents of this article.

Bibliography

Adams, Matthew, Jakob McBroome, Nicholas Maurer, Evan Pepper-Tunick, Nedda F. Saremi, Richard E. Green, Christopher Vollmers, and Russell B. Corbett-Detig. 2020a. "One Fly--One Genome: Chromosome-Scale Genome Assembly of a Single Outbred Drosophila Melanogaster." *Nucleic Acids Research* 48 (13): e75–e75.

Adams, M. D. 2000. "The Genome Sequence of Drosophila Melanogaster." *Science*. https://doi.org/10.1126/science.287.5461.2185.

AlHaj Abed, Jumana, Jelena Erceg, Anton Goloborodko, Son C. Nguyen, Ruth B. McCole, Wren Saylor, Geoffrey Fudenberg, et al. 2019. "Highly Structured Homolog Pairing Reflects Functional Organization of the Drosophila Genome." *Nature Communications* 10 (1): 4485.

Ali, Siraj M., Thomas Hensing, Alexa B. Schrock, Justin Allen, Eric Sanford, Kyle Gowen, Atul Kulkarni, 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." *The Oncologist* 21 (6): 762–70.

Al'Khafaji, Aziz M., Jonathan T. Smith, Kiran V. Garimella, Mehrtash Babadi, Moshe Sade-Feldman, Michael Gatzen, Siranush Sarkizova, et al. 2021. "High-Throughput RNA Isoform Sequencing Using Programmable cDNA Concatenation." *bioRxiv*. https://doi.org/10.1101/2021.10.01.462818.

Ashton, David T., Peter A. Ritchie, and Maren Wellenreuther. 2017. “Fifteen Years of Quantitative Trait Loci Studies in Fish: Challenges and Future Directions.” *Molecular Ecology* 26 (6): 1465–76.

Barciszewska, Mirosława Z., Patrick M. Perrigue, and Jan Barciszewski. 2016. “tRNA--the Golden Standard in Molecular Biology.” *Molecular bioSystems* 12 (1): 12–17.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. “High-Resolution Profiling of Histone Methylations in the Human Genome.” *Cell* 129 (4): 823–37.

Baslan, Timour, Sam Kovaka, Fritz J. Sedlazeck, Yanming Zhang, Robert Wappel, Sha Tian, Scott W. Lowe, Sara Goodwin, and Michael C. Schatz. 2021. “High Resolution Copy Number Inference in Cancer Using Short-Molecule Nanopore Sequencing.” *Nucleic Acids Research*, September. https://doi.org/10.1093/nar/gkab812.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. “Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes.” *Methods* 58 (3): 268–76.

Blow, Frances, and Angela E. Douglas. 2019. “The Hemolymph Microbiome of Insects.” *Journal of Insect Physiology* 115 (May): 33–39.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18.

Burton, Joshua N., Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. 2013. “Chromosome-Scale Scaffolding of de Novo Genome Assemblies Based on Chromatin Interactions.” *Nature Biotechnology* 31 (12): 1119–25.

Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017. “BBMerge - Accurate Paired Shotgun Read Merging via Overlap.” *PloS One* 12 (10): e0185056.

Byrne, Ashley, Charles Cole, Roger Volden, and Christopher Vollmers. 2019. “Realizing the Potential of Full-Length Transcriptome Sequencing.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 374 (1786): 20190097.

Byrne, Ashley, Megan A. Supple, Roger Volden, Kristin L. Laidre, Beth Shapiro, and Christopher Vollmers. 2019. “Depletion of Hemoglobin Transcripts and Long-Read Sequencing Improves the Transcriptome Annotation of the Polar Bear (Ursus Maritimus).” *Frontiers in Genetics* 10 (July): 643.

Chakraborty, Mahul, James G. Baldwin-Brown, Anthony D. Long, and J. J. Emerson. 2016. “Contiguous and Accurate de Novo Assembly of Metazoan Genomes with Modest Long Read Coverage.” *Nucleic Acids Research* 44 (19): e147.

Chakraborty, Mahul, Nicholas W. VanKuren, Roy Zhao, Xinwen Zhang, Shannon Kalsow, and J. J. Emerson. 2018. “Hidden Genetic Variation Shapes the Structure of Functional Elements in Drosophila.” *Nature Genetics* 50 (1): 20–25.

Chapman, Jarrod A., Isaac Ho, Sirisha Sunkara, Shujun Luo, Gary P. Schroth, and Daniel S. Rokhsar. 2011. “Meraculous: De Novo Genome Assembly with Short Paired-End Reads.” *PloS One* 6 (8): e23501.

Cole, Charles, Ashley Byrne, Matthew Adams, Roger Volden, and Christopher Vollmers. 2020. “Complete Characterization of the Human Immune Cell Transcriptome Using Accurate Full-Length cDNA Sequencing.” *Genome Research* 30 (4): 589–601.

Cooper, K. W. 1948. “The Evidence for Long Range Specific Attractive Forces during the Somatic Pairing of Dipteran Chromosomes.” *The Journal of Experimental Zoology* 108 (3): 327–35.

Corbett-Detig, Russell B., Iskander Said, Maria Calzetta, Max Genetti, Jakob McBroome, Nicholas W. Maurer, Vincenzo Petrarca, Alessandra Della Torre, and Nora J. Besansky. 2019. “Fine-Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the Pecies Complex Using Proximity-Ligation Sequencing.” *Genetics*, October. https://doi.org/10.1534/genetics.119.302385.

Corbett-Detig, Russell, and Rasmus Nielsen. 2017. “A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy.” *PLoS Genetics* 13 (1): e1006529.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. "De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science* 356 (6333): 92–95.

Edge, Peter, Vineet Bafna, and Vikas Bansal. 2017. "HapCUT2: Robust and Accurate Haplotype Assembly for Diverse Sequencing Technologies." *Genome Research* 27 (5): 801–12.

ENCODE Project Consortium. 2004. "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science* 306 (5696): 636–40.

Eppig, Janan T., Joel E. Richardson, James A. Kadin, Martin Ringwald, Judith A. Blake, and Carol J. Bult. 2015. "Mouse Genome Informatics (MGI): Reflecting on 25 Years." *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 26 (7-8): 272–84.

Frankish, Adam, Sílvia Carbonell-Sala, Mark Diekhans, Irwin Jungreis, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, et al. 2023. "GENCODE: Reference Annotation for the Human and Mouse Genomes in 2023." *Nucleic Acids Research* 51 (D1): D942–49.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Glinos, Dafni A., Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, et al. 2022. “Transcriptome Variation in Human Tissues Revealed by Long-Read Sequencing.” *Nature* 608 (7922): 353–59.

Gohl, Daryl M., Alessandro Magli, John Garbe, Aaron Becker, Darrell M. Johnson, Shea Anderson, Benjamin Auch, et al. 2019. “Measuring Sequencer Size Bias Using REcount: A Novel Method for Highly Accurate Illumina Sequencing-Based Quantification.” *Genome Biology* 20 (1): 85.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40 (Database issue): D1178–86.

Grenier, Jennifer K., J. Roman Arguello, Margarida Cardoso Moreira, Srikanth Gottipati, Jaaved Mohammed, Sean R. Hackett, Rachel Boughton, Anthony J. Greenberg, and Andrew G. Clark. 2015. “Global Diversity Lines - a Five-Continent Reference Panel of Sequenced Drosophila Melanogaster Strains.” *G3* 5 (4): 593–603.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29 (8): 1072–75.

Gu, Wei, Xianding Deng, Marco Lee, Yasemin D. Sucu, Shaun Arevalo, Doug Stryke, Scot Federman, et al. 2021. "Rapid Pathogen Detection by Metagenomic next-Generation Sequencing of Infected Body Fluids." *Nature Medicine* 27 (1): 115–24.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62.

Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107 (1): 1–8.

Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.

Kaessmann, Henrik. 2010. "Origins, Evolution, and Phenotypic Impact of New Genes." *Genome Research* 20 (10): 1313–26.

Kawai, J., A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, et al. 2001. "Functional Annotation of a Full-Length Mouse cDNA Collection." *Nature* 409 (6821): 685–90.

Kingan, Sarah B., Haynes Heaton, Juliana Cudini, Christine C. Lambert, Primo Baybayan, Brendan D. Galvin, Richard Durbin, Jonas Korlach, and Mara K. N. Lawniczak. 2019. "A High-Quality De Novo Genome Assembly from a Single Mosquito Using PacBio Sequencing." *Genes* 10 (1). https://doi.org/10.3390/genes10010062.

Kingan, Sarah B., Julie Urban, Christine C. Lambert, Primo Baybayan, Anna K. Childers, Brad Coates, Brian Scheffler, Kevin Hackett, Jonas Korlach, and Scott M. Geib. 2019. "A High-Quality Genome Assembly from a Single, Field-Collected Spotted Lanternfly (Lycorma Delicatula) Using the PacBio Sequel II System." *GigaScience* 8 (10). https://doi.org/10.1093/gigascience/giz122.

Kuderna, Lukas F. K., Esther Lizano, Eva Julià, Jessica Gomez-Garrido, Aitor Serres-Armero, Martin Kuhlwilm, Regina Antoni Alandes, et al. 2019. "Selective Single Molecule Sequencing and Assembly of a Human Y Chromosome of African Origin." *Nature Communications* 10 (1): 4.

Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. "Versatile and Open Software for Comparing Large Genomes." *Genome Biology* 5 (2): R12.

Lack, Justin B., Charis M. Cardeno, Marc W. Crepeau, William Taylor, Russell B. Corbett-Detig, Kristian A. Stevens, Charles H. Langley, and John E. Pool. 2015. "The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila Melanogaster Genomes, Including 197 from a Single Ancestral Range Population." *Genetics* 199 (4): 1229–41.

LaFave, Matthew C., and Shawn M. Burgess. 2014. *sam2pairwise Version 1.0.0*. https://doi.org/10.5281/zenodo.11377.

Li, Chenhao, Kern Rei Chng, Esther Jia Hui Boey, Amanda Hui Qi Ng, Andreas Wilm, and Niranjan Nagarajan. 2016. "INC-Seq: Accurate Single Molecule Reads Using Nanopore Sequencing." *GigaScience* 5 (1): 34.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.

Li, F., X. Zhao, M. Li, K. He, C. Huang, Y. Zhou, Z. Li, and J. R. Walters. 2019. "Insect Genomes: Progress and Challenges." *Insect Molecular Biology* 28 (6): 739–58.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *arXiv Preprint arXiv:1303. 3997*. https://arxiv.org/abs/1303.3997.

Li, Heng. 2016. "Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences." *Bioinformatics* 32 (14): 2103–10.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, and Richard Durbin. 2011. “Inference of Human Population History from Individual Whole-Genome Sequences.” *Nature*. https://doi.org/10.1038/nature10231.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Maples, Brian K., Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. 2013. “RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.” *American Journal of Human Genetics* 93 (2): 278–88.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.

McGarvey, Kelly M., Tamara Goldfarb, Eric Cox, Catherine M. Farrell, Tripti Gupta, Vinita S. Joardar, Vamsi K. Kodali, et al. 2015. “Mouse Genome Annotation by the RefSeq Project.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 26 (9-10): 379–90.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

Medina, Paloma, Shelbi L. Russell, and Russell Corbett-Detig. n.d. “Deep Data Mining Reveals Variable Abundance and Distribution of Microbial Reproductive Manipulators within and among Diverse Host Species.” https://doi.org/10.1101/679837.

Mitchell, Patrick K., Leyi Wang, Bryce J. Stanhope, Brittany D. Cronk, Renee Anderson, Shipra Mohan, Lijuan Zhou, et al. 2022. “Multi-Laboratory Evaluation of the Illumina iSeq Platform for Whole Genome Sequencing of Salmonella, Escherichia Coli and Listeria.” *Microbial Genomics* 8 (2). https://doi.org/10.1099/mgen.0.000717.

Morillon, Antonin, and Daniel Gautheret. 2019. “Bridging the Gap between Reference and Real Transcriptomes.” *Genome Biology* 20 (1): 112.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. “Initial Sequencing and Comparative Analysis of the Mouse Genome.” *Nature* 420 (6915): 520–62.

Munro, Rory, Roberto Santos, Alexander Payne, Teri Forey, Solomon Osei, Nadine Holmes, and Matt Loose. 2021. “MinoTour, Real-Time Monitoring and Analysis for Nanopore Sequencers.” *bioRxiv*. https://doi.org/10.1101/2021.09.10.459783.

Navarro Gonzalez, Jairo, Ann S. Zweig, Matthew L. Speir, Daniel Schmelter, Kate R. Rosenbloom, Brian J. Raney, Conner C. Powell, et al. 2021. “The UCSC Genome Browser Database: 2021 Update.” *Nucleic Acids Research* 49 (D1): D1046–57.

Newman, Aaron M., Scott V. Bratman, Jacqueline To, Jacob F. Wynne, Neville C. W. Eclov, Leslie A. Modlin, Chih Long Liu, et al. 2014. “An Ultrasensitive Method for Quantitating Circulating Tumor DNA with Broad Patient Coverage.” *Nature Medicine* 20 (5): 548–54.

Nohr, Erik, Christian A. Kunder, Carol Jones, Shirley Sutton, Eula Fung, Hongbo Zhu, Sharon J. Feng, 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*. https://doi.org/10.1101/870634.

Oki, Shinya, Tazro Ohta, Go Shioi, Hideki Hatanaka, Osamu Ogasawara, Yoshihiro Okuda, Hideya Kawaji, Ryo Nakaki, Jun Sese, and Chikara Meno. 2018. “ChIP-Atlas: A Data-Mining Suite Powered by Full Integration of Public ChIP-Seq Data.” *EMBO Reports* 19 (12). https://doi.org/10.15252/embr.201846255.

Oliphant, Travis E. 2007. “Python for Scientific Computing.” *Computing in Science & Engineering* 9 (3): 10–20.

Pardo-Palacios, Francisco J., Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, et al. 2022. “Systematic Assessment of Long-Read RNA-Seq Methods for Transcript Identification and Quantification.” https://doi.org/10.6084/m9.figshare.19642383.v1.

Pardo-Palacios, Francisco, Fairlie Reese, Silvia Carbonell-Sala, Mark Diekhans, Cindy Liang, Dingjie Wang, Brian Williams, et al. 2021. “Systematic Assessment of Long-Read RNA-Seq Methods for Transcript Identification and Quantification,” August. https://doi.org/10.21203/rs.3.rs-777702/v1.

Pietri, Jose E., Heather DeBruhl, and William Sullivan. 2016. “The Rich Somatic Life of Wolbachia.” *MicrobiologyOpen* 5 (6): 923–36.

Poplin, Ryan, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, et al. 2018. “A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks.” *Nature Biotechnology* 36 (10): 983–87.

Putnam, Nicholas H., Brendan L. O’Connell, Jonathan C. Stites, Brandon J. Rice, Marco Blanchette, Robert Calef, Christopher J. Troll, et al. 2016. “Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage.” *Genome Research* 26 (3): 342–50.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.

Ricci, William A., Laura Levin, and Xiaoyu Zhang. 2020. “Genome-Wide Profiling of Histone Modifications with ChIP-Seq.” *Methods in Molecular Biology* 2072: 101–17.

Rice, Edward S., and Richard E. Green. 2019. “New Approaches for Genome Assembly and Scaffolding.” *Annual Review of Animal Biosciences* 7 (February): 17–40.

Ruan, Jue, and Heng Li. 2019. “Fast and Accurate Long-Read Assembly with wtdbg2.” *bioRxiv*. https://doi.org/10.1101/530972.

Russell, Shelbi L., Laura Chappell, and William Sullivan. 2019. “A Symbiont’s Guide to the Germline.” *Current Topics in Developmental Biology* 135 (May): 315–51.

Shafin, Kishwar, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, et al. 2021. “Haplotype-Aware Variant Calling Enables High Accuracy in Nanopore Long-Reads Using Deep Neural Networks.” *bioRxiv*. https://doi.org/10.1101/2021.03.04.433952.

Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. “A Single-Molecule Long-Read Survey of the Human Transcriptome.” *Nature Biotechnology* 31 (11): 1009–14.

Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. “DNA Sequencing at 40: Past, Present and Future.” *Nature* 550 (7676): 345–53.

Silvestre-Ryan, Jordi, and Ian Holmes. 2021. “Pair Consensus Decoding Improves Accuracy of Neural Network Basecallers for Nanopore Sequencing.” *Genome Biology* 22 (1): 38.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12.

Steijger, Tamara, Josep F. Abril, Pär G. Engström, Felix Kokocinski, RGASP Consortium, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. 2013. “Assessment of Transcript Reconstruction Methods for RNA-Seq.” *Nature Methods* 10 (12): 1177–84.

Stein, Lincoln D. 2010. “The Case for Cloud Computing in Genome Informatics.” *Genome Biology* 11 (5): 207.

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. “Simultaneous Epitope and Transcriptome Measurement in Single Cells.” *Nature Methods* 14 (9): 865–68.

Stork, Nigel E. 2018. “How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?” *Annual Review of Entomology* 63 (January): 31–45.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector Del Risco, Marc Ferrell, et al. 2018. “SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification.” *Genome Research*, February. https://doi.org/10.1101/gr.222976.117.

Thirunavukarasu, Deepak, Lauren Y. Cheng, Ping Song, Sherry X. Chen, Mitesh J. Borad, Lawrence Kwong, Phillip James, Daniel J. Turner, and David Yu Zhang. 2021. “Oncogene Concatenated Enriched Amplicon Nanopore Sequencing for Rapid, Accurate, and Affordable Somatic Mutation Detection.” *Genome Biology* 22 (1): 227.

Treangen, Todd J., and Steven L. Salzberg. 2011. “Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions.” *Nature Reviews. Genetics* 13 (1): 36–46.

Valliyodan, B., S. B. Cannon, P. E. Bayer, and S. Shu. 2019. “Construction and Comparison of Three Reference‐quality Genome Assemblies for Soybean.” *The Plant*. https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.14500.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72.

Volden, Roger, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. 2018. “Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell cDNA.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (39): 9726–31.

Volden, Roger, Kayla Schimke, Ashley Byrne, Danilo Dubocanin, Matthew Adams, and Christopher Vollmers. 2022. “Identifying and Quantifying Isoforms from Accurate Full-Length Transcriptome Sequencing Reads with Mandalorion.” *bioRxiv*. https://doi.org/10.1101/2022.06.29.498139.

Volden, Roger, and Christopher Vollmers. 2020. “Highly Multiplexed Single-Cell Full-Length cDNA Sequencing of Human Immune Cells with 10X Genomics and R2C2.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.01.10.902361.

Vollmers, Apple Cortez, Honey E. Mekonen, Sophia Campos, Susan Carpenter, and Christopher Vollmers. 2021a. “Generation of an Isoform-Level Transcriptome Atlas of Macrophage Activation.” *Journal of Biological Chemistry*. https://doi.org/10.1016/j.jbc.2021.100784.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *PloS One* 9 (11): e112963.

Wilson, Brandon D., Michael Eisenstein, and H. Tom Soh. 2019. “High-Fidelity Nanopore Sequencing of Ultra-Short DNA Targets.” *Analytical Chemistry* 91 (10): 6783–89.

Worley, Kim C., Stephen Richards, and Jeffrey Rogers. 2017. “The Value of New Genome References.” *Experimental Cell Research* 358 (2): 433–38.

Yuan, Yuxuan, Philipp E. Bayer, Jacqueline Batley, and David Edwards. 2017. “Improvements in Genomic Technologies: Application to Crop Genomics.” *Trends in Biotechnology* 35 (6): 547–58.

Zee, Alexander, Dori Z. Q. Deng, Matthew Adams, Kayla D. Schimke, Russell Corbett-Detig, Shelbi L. Russell, Xuan Zhang, Robert J. Schmitz, and Christopher Vollmers. 2022. “Sequencing Illumina Libraries at High Accuracy on the ONT MinION Using R2C2.” *Genome Research* 32 (11-12): 2092–2106.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137.