Optimizing experimental design for genome sequencing and assembly with Oxford Nanopore Technologies

John M. Sutton[1] and Janna L. Fierst[1]*

Summary (150 words)

High quality reference genome sequences are the core of modern genomics. Oxford Nanopore Technologies (ONT) produces inexpensive DNA sequences in excess of 100,000 nucleotides but error rates remain >10% and assembling these sequences, particularly for eukaryotes, is a non-trivial problem. To date there has been no comprehensive attempt to generate experimental design for ONT genome sequencing and assembly. Here, we simulate ONT and Illumina DNA sequence reads for *Escherichia coli*, *Caenorhabditis elegans*, *Arabidopsis thaliana,* and *Drosophila melanogaster*. We quantify the influence of sequencing coverage, assembly software and experimental design on *de novo* genome assembly and error correction to predict the optimum sequencing strategy for these organisms. We show proof of concept using real ONT data generated for the nematode *Caenorhabditis remanei*. ONT sequencing is inexpensive and accessible, and our quantitative results will be helpful for a broad array of researchers seeking guidance for *de novo* genome assembly projects.

[1] Department of Biological Sciences; The University of Alabama; Tuscaloosa, AL 35487-0344; USA
*Correspondence: janna.l.fierst@ua.edu

Introduction

The ability to sequence molecular fragments has created an entirely new field of biology, genomics. In 1951, Frederick Sanger first sequenced amino acids (Sanger and Tuppy, 1951a, Sanger and Tuppy, 1951b); in 1964 Robert Holley sequenced RNA and extensions of these works led to DNA sequencing being possible (Holley et al., 1965). The first forms of DNA sequencing would follow in the Wu Lab at Cornell University in 1970 (Wu and Taylor, 1971). Wu's methods were then expanded upon by Sanger in the mid 1970's (Sanger et al., 1973) and later commercialized making sequencing technology available for scientific discovery. The advent of Illumina's high-throughput sequencing-by-synthesis technology resulted in next-generation sequencing and opened the door for rapid expansion of the genomics field (Zhang et al., 2011).

The higher accuracy of Illumina data is essential for single nucleotide polymorphism (SNP) detection or other fine-scale analyses, but the short read-length (between 50-250 nucleotides) is a challenge for genome assembly algorithms and detecting structural variants. The third generation of sequencing focuses on long sequence reads (>10,000 nucleotides) and reading nucleotides from single molecules. Currently available long-read sequencing technologies prioritize read length at the expense of accuracy. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the current front-runners in long-read sequencing platforms; both are capable of average read lengths in the tens of thousands of base pairs and, theoretically, entire chromosomes can be sequenced in a single read (Quail et al., 2012), (Schneider and Dekker, 2012) platforms and parameters are given in Table 1).

Many factors affect the *de novo* genome assembly. Genome size increases the size of the "puzzle" to put together, while the size of the pieces (sequences) remains the same. Polyploidy

46 can create scenarios where many sites in a genome look highly similar to one another, making it
47 tough to place these regions within an assembled genome (Kyriakidou et al., 2018, Claros et al.,
48 2012). For non-haploid organisms, there is a potential for heterozygosity and at these sites the
49 effective sequencing coverage is cut in half.

50 Repetitive regions, mobile genetic elements, and diversity between individuals in a
51 population also create unique challenges. Repetitive regions introduce a major hurdle for
52 Illumina data (Treangen and Salzberg, 2011) as repeats longer than the maximum read length
53 (often 150 bp) cannot be properly placed by the assembler, thus creating a break in the assembly.
54 Often these repeat regions are present in more than one location in the genome and without
55 contextual information it is difficult to identify how many copies exist in the complete genome.
56 For example, *Alu* repeat elements reach >1 million copies in the human genome (Consortium,
57 2001). With repeats making up a significant portion of larger genomes, it is possible to over- or
58 under-assemble in *de novo* genome sequences.

59 Individual diversity in a population plays a key role when pooled data must be used. This
60 is often encountered when working with small, non-clonal metazoans where the necessary
61 amount of DNA cannot be acquired from a single individual. This pooled-data compounds the
62 issues of ploidy and heterozygosity.

63 The read sizes of PacBio or ONT data can theoretically solve these problems. The long
64 sequence reads span repetitive regions, potentially allowing for the identification of the exact
65 size and location of these repeats on a chromosome. Long sequence reads increase the puzzle
66 piece size for assembly and require less sequencing effort to span the entire genome. In fact, with
67 microbial genomes, it is possible to assemble highly accurate, complete genomes with just long-
68 read sequence data (Koren et al., 2013). Since Illumina short-read data are orders of magnitude

69 more accurate than their long-read counterparts, software packages such as Pilon (Walker et al.,
70 2014) use Illumina sequences for error correction or 'polishing'. Using both short- and long-read
71 sequencing together can overcome the short-comings of both to create higher quality genome
72 assemblies (Zimin et al., 2017).

73 ONT offers several advantages over PacBio. Nanopore sequencing relies on running
74 molecular fragments through engineered nanopores and recording the resulting alterations in
75 electrical current. The technology is versatile and can be used for DNA sequencing, mRNA
76 sequencing, amplification-free mRNA quantification (Byrne et al., 2017), and measuring DNA
77 base modifications like methylation (Jain et al., 2016, Simpson et al., 2017). ONT is relatively
78 inexpensive: for a similar cost, 10-100x the amount of sequence can be generated with ONT
79 when compared with PacBio. ONT libraries can be readily prepared with low amounts of input
80 DNA, an important consideration when studying organisms that are small or difficult to sample.
81 ONT currently offers two inexpensive platforms, the MinION and GridION, that are designed to
82 be used in individual research laboratories. The MinION is a portable sequencing device that can
83 attach to a standard computer via USB. The GridION has 5x the sequencing capacity of a
84 MinION and is designed for high computational requirements. For these reasons we have chosen
85 to study ONT and quantify how this inexpensive, accessible technology may be best utilized to
86 produce high-quality assembled reference genome sequences. A similar project analyzed
87 experimental design for PacBio (Chakraborty et al., 2016) but there are currently few
88 experimental design guidelines for ONT sequencing and assembly.

89 Despite advantages in cost and accessibility, ONT sequence reads are uniquely
90 challenging for genome sequence assembly. DNA molecules do not move through protein
91 nanopores at a constant rate and changes in current are a composite signature reflecting 3-5

92 nucleotides occupying the nanopore (for R9.4.1 flow cells). The signal processing has trouble
93 detecting changes in current with homopolymers >5 nucleotides (single nucleotide repeats, for
94 example AAAAAA) and can truncate these regions. As a result, ONT sequence reads contain
95 nucleotides that have been incorrectly identified, inserted and deleted, (Fig. 1A, C). This error
96 structure results in relatively few large contiguous stretches of correctly identified nucleotides
97 (Fig. 1B, D) and is uniquely challenging for assembling genome sequences.

98 In this article we simulate DNA sequence read sets for the model organisms *Escherichia*
99 *coli*, *Caenorhabditis elegans*, *Arabidopsis thaliana,* and *Drosophila melanogaster*. For each
100 organism we assemble ONT sequence read sets at different sequencing depths, or the average
101 times a nucleotide in the genome is sequenced in our simulated library. We measure the
102 contiguity, completeness, and accuracy of each assembled sequence relative to the current
103 reference genome. Many *de novo* assembly projects target organisms without reference genomes
104 and we also measure the conservation of a set of genes thought to be found in single copy in each
105 organism. We use these measures to identify the optimal depth and strategy of sequencing
106 required for competent assembly.

107 Pure ONT datasets result in superior assembled genome sequences but can be cost-
108 prohibitive and inaccurate. We also analyze assembled sequences from 'hybrid' DNA read
109 datasets composed of ONT and Illumina sequence reads. Error correcting or 'polishing' ONT
110 assembled sequences with higher-accuracy Illumina reads is a necessary step in producing a
111 genome sequence. We sequence and assemble an empirical ONT-generated genome sequence for
112 the nematode *C. remanei* and measure the effects of polishing the assembled sequence with
113 varying depths of Illumina sequence reads.

We find that ONT approaches can produce highly contiguous genome assemblies with relatively high sequencing coverage of >100x, 5x higher than the current recommendations. Pure ONT sequencing and assembly outperforms our tested hybrid approach. For organisms where >100x ONT coverage cannot be generated, we find the success of hybrid assembly is determined by the sequencing coverage of the Illumina data. We also find that the use of Illumina data, even at low 20x sequencing depths, increases accuracy through iterative polishing.

Results

We simulated 150bp paired-end Illumina DNA libraries with the software ART (Huang et al., 2012) and ONT DNA libraries with the software NanoSim (Yang et al., 2017). Both software programs utilize an assembled sequence to generate simulated libraries with read profiles similar to an empirically generated DNA library. NanoSim additionally requires real ONT flowcell data to simulate the unique, organism-specific ONT mismatch, insertion and deletion rates.

We assembled ONT libraries with the Canu software package (Koren et al., 2017) and 'hybrid' ONT and Illumina libraries with the MaSuRCA software package (Zimin et al., 2017). We measured genome statistics relative to the reference sequence of each organism with QUAST (Gurevich et al., 2013). We assessed contiguity and accuracy of the assembled sequence through eight statistics:

(1) **NG50** is a size median statistic and indicates that 50% of the expected assembled genome sequence (where the expectation is based on a known reference) is contained in contiguous sequences that large or larger.

135 (2) **NGA50** is a similar size median but indicates that 50% of the expected assembled genome
136 sequence that aligns to the reference genome is contained in contiguous sequences that large or
137 larger.
138 (3) **LG50** is the number of linkage groups or contiguous assembled sequences containing 50% of
139 the expected assembled genome sequence.
140 (4) **LGA50** is similar but again measured in the portion of the assembled sequence that aligns to
141 the reference genome.
142 (5) **Genome fraction (%)** is the fraction of reference genome captured in the assembled
143 sequence.
144 (6) **Duplication** measures the fraction of reference genome found multiple times in the
145 assembled sequence.
146 (7) **Mismatches** is the number of mismatched nucleotides per 100,000 nucleotides or base pairs
147 (bp).
148 (8) **Indels** are the number of insertions and/or deletions per 100,000 nucleotides.
149 For our metazoan organisms we also used the software package BUSCO to search for a set of
150 unique genes that are expected to be conserved in single copy in an evolutionarily related group
151 of organisms (Simão et al., 2018).
152
153 ***Escherichia coli***
154 Canu assemblies (Koren et al., 2017) resulted in single contigs that could be circularized when
155 using high levels of coverage. The number of assembled contigs decreased with increasing
156 coverage levels (Fig. 2A, B). The best overall performance was produced by a dataset containing
157 50,000 reads (~62x coverage or ~290Mbp). This produces a single circular contig that matched

158 the reference genome 100% and had a duplication ratio of just 1.001 (1 duplicated base per
159 1000bp). The worst Canu assembly used 15,000 reads (~19x coverage) and produced 6 total
160 contigs.

161 The hybrid assembly approach using MaSuRCA (Zimin et al., 2017) performed well for
162 *E. coli*. Two of the tested hybrid sets were able to assemble the genome into a single contig.
163 These two sets had differing coverage depths for both long- and short-reads (Fig. 2A, B). Here
164 we also noted that MaSuRCA consistently was unable to perform as well as Canu (Koren et al.,
165 2017) when given the same long-read dataset. For instance, the top performing Canu run used
166 50,000 ONT reads (~62x coverage); the same reads passed through MaSuRCA produced 2
167 contigs, regardless of coverage from Illumina data (Fig. 2B).

168 We used MaSuRCA to assemble a set of paired end Illumina sequences (DNA reads from
169 other side of a single molecule) to test the influence of adding ONT data to the assembly process.
170 The resulting sequence had a genome fraction of 98.76% and was fragmented into 74 contiguous
171 pieces. The accuracy, 1.7 mismatches and 0.04 indels per 100kbp, was higher than many of the
172 hybrid assembled sequences. This indicates that the inclusion of ONT data can introduce
173 misassemblies in hybrid approaches.

174

175 ***Caenorhabditis elegans***

176 The ONT assemblies for *C. elegans* show a general improvement in contiguity as coverage depth
177 increases (Fig. 3A,B). However, it does take more overall coverage to approach a chromosome-
178 level assembly. This is likely due to increased genome size, increased genome complexity, and
179 the addition of heterozygosity in the data compared to haploid *E coli.* The top-performing
180 assembly for *C. elegans* from Canu (Koren et al., 2017) produced 14 contigs (6 chromosomes)

181 from ~336x ONT (33.6Gbp) coverage. One assembly produced fewer contigs with 264x ONT
182 coverage but had lower overall accuracy and quality scores (Fig. 3C, D). It is also worth noting
183 that smaller data sets (ones that realistically can acquired from a single ONT flow cell) still
184 produced highly contiguous assemblies with 21 or fewer contigs. We found that error correcting
185 the more contiguous assembly with Illumina paired-end sequences and the Pilon software
186 package (Walker et al., 2014) rectified the discrepancy in accuracy between the top Canu
187 performers (Table 2; S. Table 1).
188 The MaSuRCA (Zimin et al., 2017) hybrid assembly approach did not perform as well,
189 even with high ONT coverage (Fig. 3A,B). Many of the assemblies produced by MaSuRCA
190 were fragmented and much smaller than the true genome size of ~100Mbp (Fig. 3C). The
191 produced assemblies ranged from 2.26-99.6% matching with the Ensembl reference genome. We
192 assembled the Illumina paired end sequences with MaSuRCA and produced a sequence with
193 3,353 contiguous pieces totaling 99,171,998 bases in length. This sequence had relatively few
194 mismatches, insertions and deletions compared with either ONT or hybrid assemblies (Table 2).
195
196 ***Drosophila melanogaster***
197 Canu (Koren et al., 2017) assemblies of ONT data simulated from *D. melanogaster* repeated the
198 pattern seen with the *C. elegans* dataset; the most contiguous assembly was produced with 113x
199 ONT coverage (~16Gbp; Fig. 4A,B). This assembly produced 145 contiguous pieces but many
200 of these were small and the LG50 was low (Figure 4B). We identified 91.7% of the metazoan
201 genes expected to be conserved in single copy with BUSCO (Simão et al., 2018). Additional data
202 decreased contiguity with a slight increase in accuracy and increased NG50 (Fig. 4 A,D).
203 Following error correction with Pilon (Walker et al., 2014), the 113x coverage assembly had

204 95.5% of the expected single copy metazoan genes (Simão et al., 2018) found in single,
205 duplicated or fragmented copy with 94% in single copy (Fig. 5A). In comparison, the *D.*
206 *melanogaster* reference sequence contains 95.6% of the expected single copy metazoan genes
207 with 94.1% in single copy.

208 Hybrid MaSuRCA (Zimin et al., 2017) assemblies for *D. melanogaster* performed
209 markedly worse than the pure ONT assemblies in regard to contiguity (Figure 4B). While the
210 hybrid assemblies do have an initial advantage in accuracy, this disappears after polishing with
211 Illumina sequences (Table 2). We again note that with smaller datasets, the hybrid approach
212 produced assemblies that were much smaller than the expected genome size (Fig. 4C). While not
213 as drastic as the discrepancies seen with *C. elegans,* assembly completion vs. the reference
214 ranged from 61.5%-94.5%. The top performing MaSuRCA assembly produced 220 contigs and
215 93.9% of the expected metazoan genes were identified in the sequence (Simão et al., 2018).
216

217 ***Arabidopsis thaliana***

218 The assemblies produced by Canu (Koren et al., 2017) for *A. thaliana* consistently improved
219 contiguity with increasing data. The top Canu assembly started with 420x coverage (56.7Gbp)
220 and produced 30 contiguous pieces (LG50 5 chromosomes) with 98.8% of the expected
221 Viridiplantae genes (Table 2) identified prior to polishing (Simão et al., 2018). Following
222 polishing with Pilon (Walker et al., 2014), the assembly contained 99.1% of the expected
223 Viridiplantae genes (Table 2; (Simão et al., 2018), matching that of the TAIR10 reference
224 genome for *A. thaliana*.

225 The hybrid assemblies for *A. thaliana* performed the best of the three eukaryotes. They
226 produced comparable, yet slightly less contiguous assemblies when compared the long-read only

227 approach (SFig. 1A,B). Here, we noted a similar pattern to that seen in *C. elegans* and *D.*
228 *melanogaster*, where the assembly produced is much too small in comparison to the reference
229 when using a reduced paired-end dataset (SFig. 1C). The top performing MaSuRCA (Zimin et
230 al., 2017) assembly produced 45 contiguous pieces with 98.8% of the expected Viridiplantae
231 genes identified in the sequence (Simão et al., 2018).
232
233 ***Caenorhabditis remanei***
234 We attempted to follow the trends found from the simulated data in our approach to creating a *de*
235 *novo* assembled sequence for the nematode *C. remanei*, strain PX356. *C. remanei* is an obligate
236 outcrossing species with high levels of nucleotide diversity that have hobbled previous assembly
237 attempts (Fierst et al., 2015, Barriere et al., 2009). The best assembly was achieved using 102x
238 ONT coverage (13.3Gbp) and Canu v1.9 (Koren et al., 2017). This yielded 183 contigs with
239 80.3% of the expected conserved nematode genes identified in the sequence (Simão et al., 2018).
240 A MaSuRCA-hybrid approach (Zimin et al., 2017), using 102x ONT coverage and 450x
241 paired-end coverage yielded 336 contigs and 96.6% of BUSCO single-copy genes.
242 Polishing played a large role in increasing accuracy of the real-world data. The completeness of
243 the Canu (Koren et al., 2017) assembly increased to 97.7% after Pilon (Walker et al., 2014)
244 polishing with Illumina paired end reads at ~225x average depth (Fig. 6). We tested the influence
245 of Illumina coverage on polishing and found that 97.6% of the expected conserved nematode
246 genes could be identified after polishing with just ~20x Illumina coverage (Table 3), indicating
247 that a large amount of data is not necessary to correct the majority of errors in an assembly.
248 However, this was achieved after three successive rounds of error correction with the Pilon
249 software package (Walker et al., 2014) utilizing the same Illumina DNA sequence reads. This

250 assembled sequence is less fragmented and contains a higher percentage of conserved nematode
251 genes (Fig. 6) when compared with the previously published assembly for *C. remanei* PX356
252 produced using paired-end and mate-pair data as well as a linkage map (Fierst et al., 2015).
253

254 **Discussion**

255 We have found that the Canu software package (Koren et al., 2017), using Oxford Nanopore
256 long-reads only, produces the most contiguous draft assemblies at the expense of accuracy across
257 a broad range of organisms. However, this accuracy can be improved by polishing with Illumina
258 DNA sequences and the Pilon software package (Walker et al., 2014). These contiguous
259 assemblies could only be achieved with relatively high sequencing depth, at least 100x coverage
260 across the genome. This is far higher than the current recommendations of 20x as a minimum
261 and 30-60x for ideal results. We found the discrepancy to be caused by differences in idealized
262 vs. actual read lengths. Although ONT can theoretically produce megabase-sized reads in reality
263 many of the sequence reads in 'real' projects are shorter due to handling techniques that result in
264 library fragmentation and truncated DNA molecules. During assembly the Canu software will
265 discard shorter reads and create a dataset that ideally has 40x coverage of the estimated genome
266 in reads 10,000bp or longer. Real ONT libraries, like the ones we used for simulations, have
267 many more small reads and >100x real coverage can be required to achieve effective high-
268 confidence, long read coverage. Intense effort has gone into developing robust high molecular
269 weight DNA extraction protocols that can alleviate some of these issues. However, 'real'
270 sequencing projects should aim to generate >100x coverage for reliable sequencing.

271 Overall, we found the hybrid assemblies produced by MaSuRCA (Zimin et al., 2017)
272 were fragmented with more contigs, lower NG50 values and assembled a smaller fraction of the

273 expected genome. We found, surprisingly, that given the same amount of ONT data the Canu
274 software (Koren et al., 2017) assembled a higher contiguity genome sequence when compared
275 with a hybrid MaSuRCA assembly. These hybrid assemblies contained a higher proportion of
276 expected conserved genes when compared with the raw Canu-assembled ONT read sets and had
277 fewer mismatches and insertion/deletion errors. However, Illumina sequences, even in small
278 amounts, can be used to error correct the draft assemblies produced by Canu and improve the
279 accuracy to be on par with, or better than, those produced by MaSuRCA.

280 In light of our findings, we suggest that long-read data be prioritized when undertaking
281 *de novo* genome assembly projects. Our results indicate that an assembly with ONT long reads
282 only will be the most contiguous and the inclusion of a small amount of PE data can improve
283 accuracy to high levels. For situations where this is not possible, MaSuRCA-assembled (Zimin et
284 al., 2017) Illumina and ONT read sets can produce reliable draft sequences. However, the quality
285 and contiguity of the assembled sequence is determined by Illumina read depth and effort should
286 be made to increase Illumina read depth, even if it is at the expense of ONT sequences.
287 MaSuRCA-assembled Illumina sequences have fewer mismatches and insertion/deletion errors
288 when compared with MaSuRCA-assembled ONT and Illumina hybrid read sets, indicating that
289 the inclusion of ONT sequences introduces errors. We suggest that error correction with Illumina
290 DNA sequences and the Pilon software package (Walker et al., 2014) is a necessary finishing
291 step in any assembly project utilizing ONT data.

292 Our results demonstrate that chromosome-level genome sequences are achievable with
293 sufficient ONT data. However, chromosome-level genome assemblies are often not necessary to
294 address many research questions, particularly those focused on small numbers of genes or
295 phylogenomic information. Researchers should approach genome sequencing by first

determining what genome-completion level will be sufficient for their research goals. To aid in this approach, we hope our study will help researchers determine the amount of sequencing effort, and the sequencing approaches, that will best suit their needs.

Limitations of the Study

Our study has two central limitations. First, our study is based on simulated data. Our simulated Illumina and ONT DNA sequences and resulting assembled genome sequences are limited by the quality of the available reference. Each of the model organisms we targeted has chromosome-level assemblies, but they may still have issues with contiguity and completeness. For example, they may not reach one contig per chromosome or 100% conservation of expected genes for that taxonomic division. We have made our measurements relative to the reference genome to account for this. Second, because of the time and resources necessary to assemble ONT and hybrid ONT and Illumina read sets we were not able to exhaustively search parameter space or assemble very high depth ONT read sets for our metazoan model organisms. Despite these limitations we have presented a quantitative approach to experimental design for genome sequencing and assembly that will be useful to a broad array of researchers interested in genomic questions.

Resource Availability

Lead Contact:

Janna L. Fierst: janna.l.fierst@ua.edu

Materials Availability:

This study did not generate new unique reagents.

Data and Code Availability:

E. coli genome sequence GCF_000005845.2; ONT read set SRR8154670

C. elegans genome sequence GCA_000002985.3; ONT read set ERR2092776

A. thaliana genome sequence GCA_000001735.1; ONT read set ERR2173373

D. melanogaster genome sequence GCA_000001215.4; ONT read set SRR6702603

C. remanei PX356 LFJK00000001; Illumina read set SRX3014103

Simulated genome sequences have been deposited with the Dryad data repository.

Acknowledgements

We thank Paula Adams, Denise Akob, Louis Bubrig and Joshua Millwood for helpful discussions. This work is funded by National Science Foundation uRoL grant 1921585.

Author Contributions

Conceptualization, J.M.S. and J.L.F.; Methodology, J.M.S. and J.L.F.; Formal Analysis, J.M.S. and J.L.F.; Resources, J.L.F.; Writing – Original Draft, J.M.S. and J.L.F.; Writing – Review & Editing, J.M.S. and J.L.F.; Visualization, J.L.F.; Supervision, J.L.F.

Declaration of Interests

The authors declare no competing interests.

References

Barriere, A., Yang, S. P., Pekarek, E., Thomas, C. G., Haag, E. S. & Ruvinsky, I. (2009). Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome Research* 19**,** 470-80.10.1101/gr.081851.108

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M. & Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature communications,* 8**,** 16027-16027.10.1038/ncomms16027

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research,* 44**,** e147-e147.10.1093/nar/gkw654

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P. & Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology,* 1**,** 439-459.10.3390/biology1020439

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860.doi:10.1038/35057062

Fierst, J. L., Willis, J. H., Thomas, C. G., Wang, W., Reynolds, R. M., Ahearne, T. E., Cutter, A. D. & Phillips, P. C. 2015. Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. *PLoS Genet.*10.1371/journal.pgen.1005323

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England),* 29**,** 1072-1075.10.1093/bioinformatics/btt086

Holley, R. W., Everett, G. A., Madison, J. T. & Zamir, A. (1965). Nucleotide Sequences In The Yeast Alanine Transfer Ribonucleic Acid. *J Biol Chem,* 240**,** 2122-8

Huang, W., Li, L., Myers, J. R. & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England),* 28**,** 593-594.10.1093/bioinformatics/btr708

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology,* 17**,** 239.10.1186/s13059-016-1103-0

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., McVey, S. D., Radune, D., Bergman, N. H. & Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol,* 14**,** R101.10.1186/gb-2013-14-9-r101

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.10.1101/gr.215087.116

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D. & Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in plant science,* 9**,** 1660-1660.10.3389/fpls.2018.01660

Ncbi Resource Coordinators (2017). Database Resources of the National Center for Biotechnology Information. *Nucleic acids research,* 45**,** D12-D17.10.1093/nar/gkw1071

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.10.1186/1471-2164-13-341

Sanger, F., Donelson, J. E., Coulson, A. R., Kossel, H. & Fischer, D. (1973). Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc Natl Acad Sci U S A,* 70**,** 1209-13

Sanger, F. & Tuppy, H. (1951a). The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem J,* 49**,** 463-81

Sanger, F. & Tuppy, H. (1951b). The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem J,* 49**,** 481-90

390 Schneider, G. F. & Dekker, C. (2012). DNA sequencing with nanopores. *Nature Biotechnology,* 30**,** 326-328.10.1038/nbt.2181

391 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. (2018). BUSCO: assessing genome
392 assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31**,** 3210-
393 3212.10.1093/bioinformatics/btv351

394 Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J. & Timp, W. (2017). Detecting DNA cytosine methylation
395 using nanopore sequencing. *Nature Methods,* 14**,** 407-410.10.1038/nmeth.4184

396 Treangen, T. J. & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and
397 solutions. *Nature reviews. Genetics,* 13**,** 36-46.10.1038/nrg3117

398 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S.
399 K. & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly
400 improvement. *PLoS One,* 9**,** e112963.10.1371/journal.pone.0112963

401 Wu, R. & Taylor, E. (1971). Nucleotide sequence analysis of DNA: II. Complete nucleotide sequence of the cohesive ends of
402 bacteriophage λ DNA. *Journal of Molecular Biology,* 57**,** 491-511.https://doi.org/10.1016/0022-2836(71)90105-7

403 Yang, C., Chu, J., Warren, R. L. & Birol, I. (2017). NanoSim: nanopore sequence read simulator based on statistical
404 characterization. *GigaScience,* 6**,** 1-6.10.1093/gigascience/gix010

405 Zhang, J., Chiodini, R., Badr, A. & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of*
406 *genetics and genomics = Yi chuan xue bao,* 38**,** 95-109.10.1016/j.jgg.2011.02.003

407 Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marcais, G., Yorke, J. A., Dvorak, J. & Salzberg, S. L. (2017). Hybrid
408 assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the
409 MaSuRCA mega-reads algorithm. *Genome Res,* 27**,** 787-792.10.1101/gr.213405.116

410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433

Table 1: Summary of Current DNA Sequencing Platforms

Company	Platform	Read Length	Total Output (reads per run)	Accuracy**
Illumina	MiSeq	50-600bp	1-25 million	99.9%
	Hi-Seq2500	50-500bp	300 million – 2 billion	99.9%
	NovaSeq 6000	50-150bp	32 - 40 billion	99.9%
PacBio	RSII	1,000+bp*	50-100,000	86%
	Sequel	1,000+bp* Average = 30kb	~500,000	89% (99% with HiFi)
ONT	MinION/GridION Flowcell: R9.4.1RevD	1,000+bp* Average = 8-15kb	1-5 million	~90%

* Read length for long-read technologies is highly dependent on DNA isolation and size selection protocols.
**Accuracy here is listed per base: 99.9% accuracy means there is one error per 1,000 bases sequenced.

471 **Table 2.** Genome statistics for MaSuRCA-assembled Illumina paired sequences and error corrected Canu-assembled sequences.
472 Although the Illumina paired end sequence is highly fragmented and incomplete it has fewer mismatches, insertions and deletions
473 when compared with hybrid ONT and Illumina read sets.
474

	Read Coverage		**BUSCO**					
Sample	ONT	Illumina	Single (%)	Duplicated	Fragmented	Missing	Mismatches per 100Kpb	Indels per 100Kpb
A. thaliana	0	100	98.1	0.9	0.5	0.5	10.99	1.03
	420	polished	99.1	0.5	0	0.4	11.36	7.3
C. elegans	0	200					4.78	1.02
	336	polished	97.7	0.6	0.4	1.3	0.73	0.88
D. melanogaster	0	100	93.5	1.5	0.4	4.6	10.01	1.7
	108	polished	94	1.2	0.3	4.5	8.84	12.87
E. coli	0	100	-	-	-	-	1.7	0.04
	62	polished	-	-	-	-	0.02	0.02

475
476
477
478

Table 3: Polishing the *C. remanei* assembled sequence with an Illumina library at 20x depth increases the percentage of conserved genes identified by BUSCO (Simão et al., 2018) after 3 rounds of polishing with Pilon (Walker et al., 2014). Polishing with a higher depth Illumina library (114x average sequencing coverage) produces similar results after 2 rounds of polishing with Pilon. The highest percentage of conserved genes found in single copy is achieved after 3 rounds of polishing with Pilon and a high depth Illumina library at 227x average coverage.

	Polishing		**BUSCO**			
Sample	Rounds	Coverage	Single (%)	Duplicated	Fragmented	Missing
Canu	0	20x	80.3	0.9	7.7	11.1
Assembled	1	20x	97.2	1	0.5	1.3
C.	2	20x	97.3	1	0.4	1.3
remanei	**3**	**20x**	**97.6**	**1**	**0.2**	**1.2**
	4	20x	97.6	1	0.2	1.2
	1	114x	97.6	1.1	0.3	1
	2	**114x**	**97.6**	**1**	**0.3**	**1.1**
	3	114x	97.5	1	0.3	1.2
	4	114x	97.5	1	0.3	1.2
	1	227x	97.6	1	0.3	1.1
	2	227x	97.5	1	0.3	1.1
	3	**227x**	**97.7**	**0.9**	**0.3**	**1.1**
	4	227x	97.7	0.9	0.3	1.1

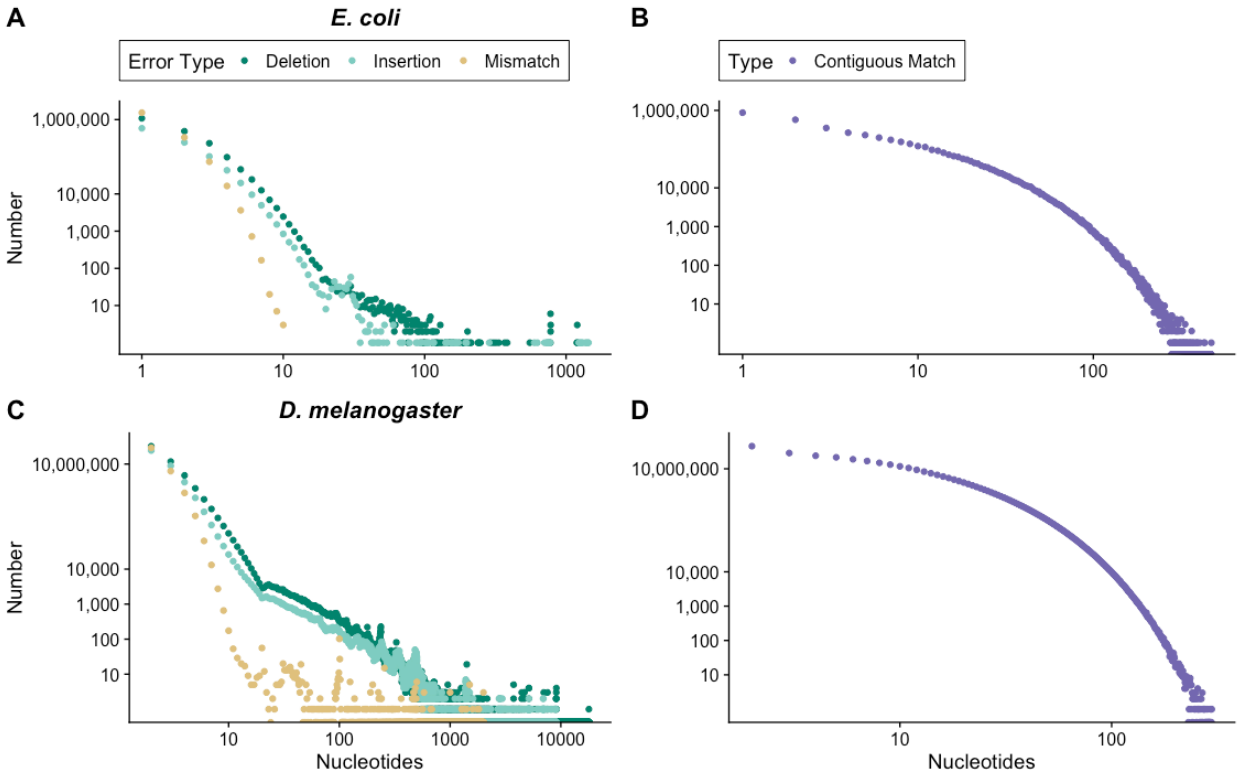


Figure 1. ONT sequence reads contain mixtures of errors including miscalled nucleotides, deletions, insertions and truncated homopolymers. When aligned to the reference genome this results in a large number of single and multi-base deletions, insertions and mismatches for (A) *E. coli* and (B) relatively few stretches of contiguous matching sequence that extend beyond a few nucleotides. For (C) *D. melanogaster* the error profile is similar but the large, repeat-rich genome results in multi-nucleotide deletions and insertions and again (D) few stretches of long, contiguous matching sequence.

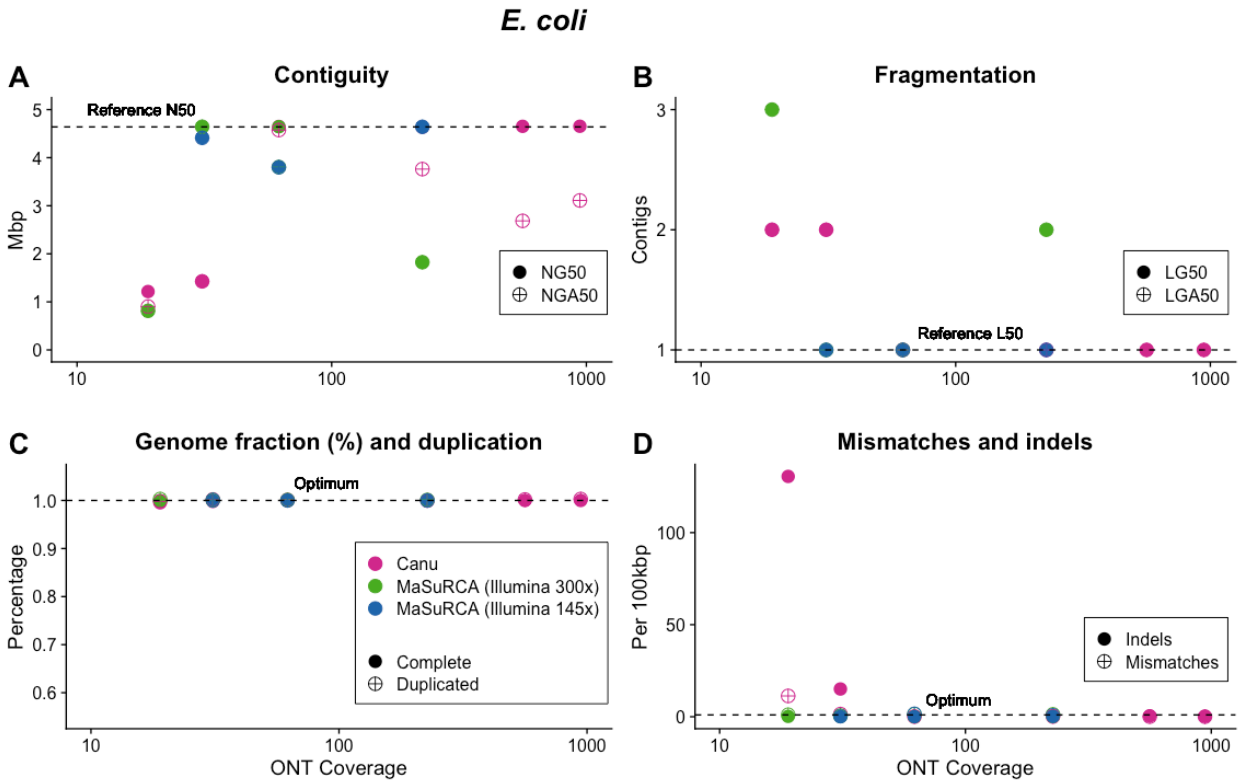


Figure 2. The *E. coli* genome is relatively small at 4.64Mbp and less complex when compared with metazoan genome sequences. Assembly of ONT libraries at relatively high coverage (>100x average sequence depth) with both Canu and MaSuRCA results in assembled sequences with (A) high contiguity; (B) low contig number converging on a single chromosome; (C) high genome completion and low duplication; and (D) few mismatches and insertion/deletion errors (indels) when compared with the reference sequence.

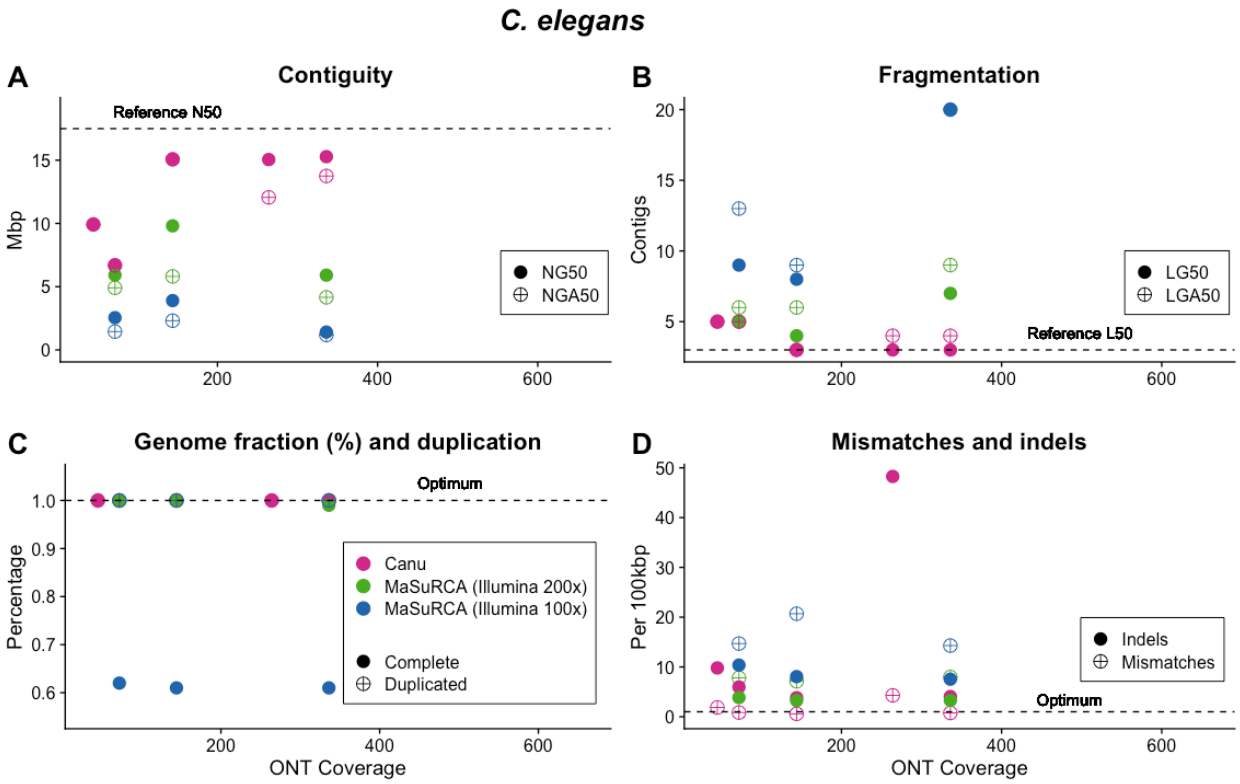


Figure 3. The diploid metazoan *C. elegans* genome is 100.8Mbp and contains complex features including introns, non-coding regions and repeat elements. Assembly of a high coverage ONT library produces a sequence with (A) high contiguity and; (B) low contig number but these statistics show a non-monotonic dependence on ONT coverage with both Canu and MaSuRCA software packages. Both the Canu assembly and a high-coverage Illumina MaSuRCA assembly had (C) a high level of completion with little duplication but; (D) but the assembled sequences had high levels of mismatches, insertions and deletions.

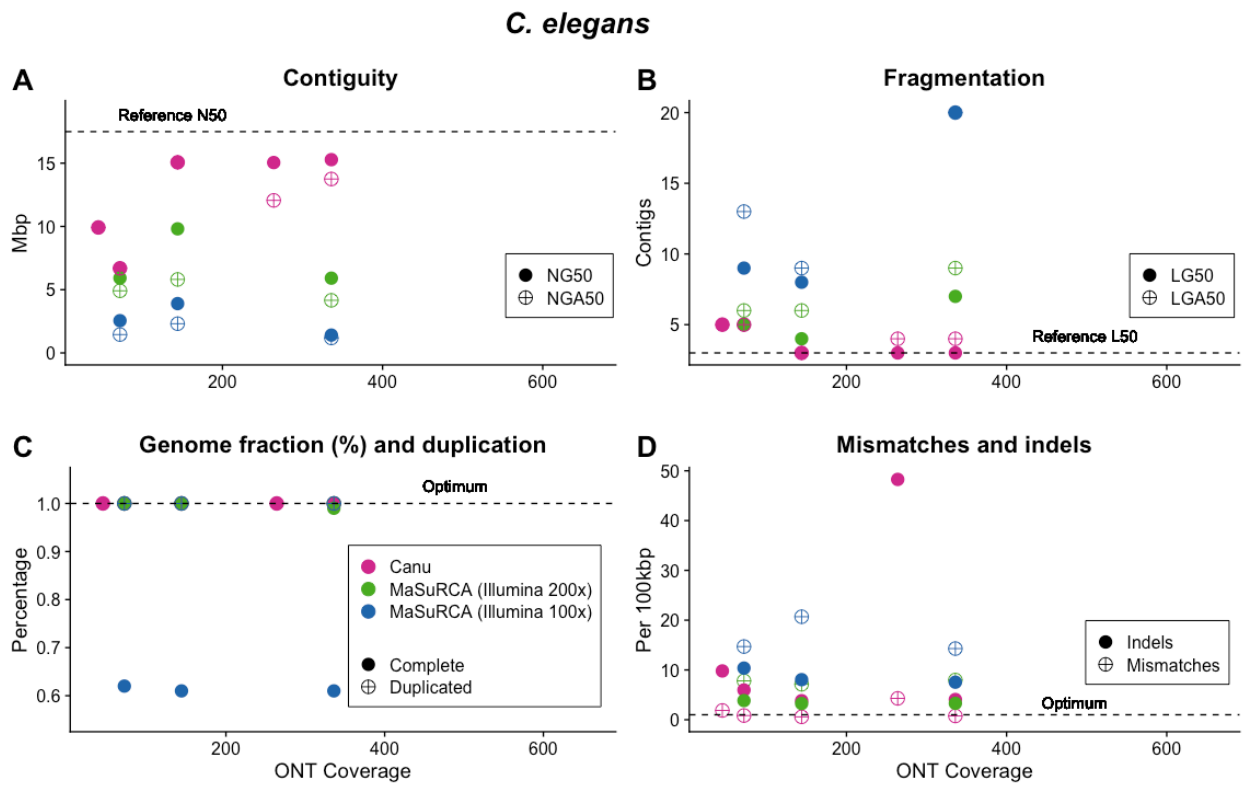


Figure 4. The *D. melanogaster* genome is 139.5Mbp and presents multiple challenges for genome sequencing and assembly including gene duplications, gene families and repetitive sequences. Assembly of a high coverage ONT library with Canu produces a sequence with (A) high contiguity; (B) low contig number; (C) a high level of completion with little duplication; (D) but the assembled sequence retains multiple mismatches and insertions and deletions (indels) when compared with the reference.

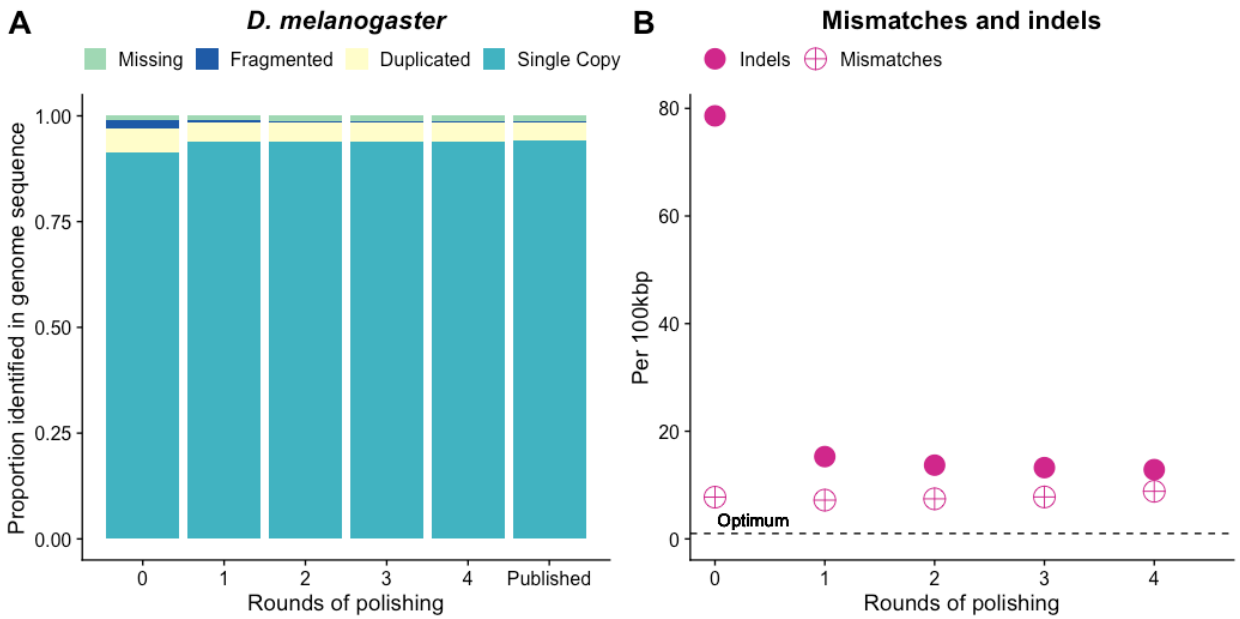


Figure 5. Polishing the assembled *D. melanogaster* sequence with Illumina libraries and the software package Pilon (Walker et al., 2014) increases (A) the number of conserved genes found in single copy; and reduces (B) the number of mismatches and indels compared with the reference sequence. The bulk of this improvement occurs after 1-2 rounds of polishing.

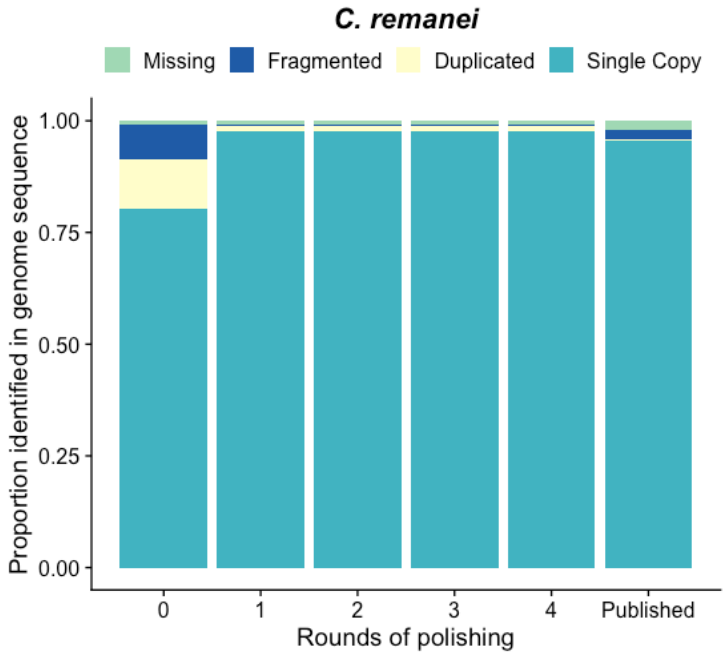


Figure 6. Polishing an empirical assembled sequence with Illumina libraries and the Pilon software package (Walker et al., 2014) increases the BUSCO completeness with a single round. The assembled sequence has a greater number of conserved genes found in single copy when compared with the published sequence (Fierst et al., 2015)

Transparent Methods

Sequence data

We obtained ONT sequences (R9 chemistry) from the National Center for Biotechnology Information (NCBI) Sequence Read Archive on February 2, 2019 (Ncbi Resource Coordinators, 2017). The *E. coli* dataset contained 11,652,330 sequenced bases in 120,151 reads (~2.5x coverage of the 4.64Mb genome), the *C. elegans* dataset contained 8,860,671,330 sequenced bases in 583,466 reads (~87.9x coverage of the 100.8Mb genome), the *A. thaliana* dataset contained 3,421,779,258 sequenced bases in 300,071 reads (~25.22x coverage of the 135.67Mb genome), and the *D. melanogaster* dataset contained sequenced bases in 663,784 reads (~32.39x coverage of the 142.57Mb genome). We obtained the reference genome sequence for *E. coli* strain K12_MG1655 from NCBI; all other reference genome sequences were obtained from Ensembl (Release 95). Accession numbers are provided under 'Resource Availability.'

Simulated libraries

We simulated 150bp paired-end Illumina DNA libraries with the software ART (version MountRainier; (Huang et al., 2012)) and ONT DNA libraries with the software NanoSim (version 2.0.0; (Yang et al., 2017)). The genome sequences of *E. coli, S. cerevisiae, C. elegans, A. thaliana,* and *D. melanogaster* were obtained from Ensembl (Release 95) and used for library simulation. ONT DNA sequencing is sensitive to organism-specific base modifications and the NanoSim (Yang et al., 2017) software requires both an assembled reference genome sequence and a set of empirically obtained Nanopore sequences for that organism. Sequence accessions are listed under 'Resource availability.'

Assembly & Polishing

Genomes were assembled using two approaches. The first used the simulated ONT read sets and Canu v1.9 (Koren et al., 2017). Each genome was assembled at decreasing coverage depths until the assembler was unable to complete an assembly with the given data. In order to minimize the influence of individual reads and stochastic assembly artifacts, each read set was generated by selecting a random subset of the full simulated dataset.

The long-read datasets that performed the best were then paired with simulated paired-end Illumina data and assembled using MaSuRCA version 3.3.9 (Zimin et al., 2017); coverage depths were adjusted for both datasets to better understand the effects of increasing or decreasing coverage on the final assembly. Here, we retained the ONT dataset to maximize our ability to draw parallels between assembly approaches. For example, the minimum ONT dataset that assembled with Canu (Koren et al., 2017) was an average of 60x coverage across the genome and this readset was used in the MaSuRCA trials with 50x and 100x Illumina coverage, respectively.

The most contiguous assemblies from the long-read only and hybrid categories were error corrected using Pilon version 1.23 (Walker et al., 2014) to determine the effect of short-read polishing on the accuracy of the draft assemblies. Each simulated assembled sequence was polished with the entire simulated paired-end data set. Four rounds of polishing were completed for each assembly with statistics measured after each round with QUAST (Gurevich et al., 2013) and BUSCO (Simão et al., 2018).

Evaluation

We used the software BUSCO version 4.0.1 (Simão et al., 2018) to identify conserved gene sets. Briefly, BUSCO searches assembled DNA sequences for a set of unique genes that are expected to be conserved in single copy in an evolutionarily related group of organisms. We used Nematoda_odb10 for *C. elegans*, Metazoan_odb10 for *D. melanogaster*, and Viridiplantae_odb10 for *A. thaliana*. We also measured BUSCO completeness with the Diptera odb10 for the *D. melanogaster* assemblies but found that in some instances our assembled sequence contained a greater proportion of conserved genes than the reference sequence. We chose to focus on the Metazoan_odb10 for *D. melanogaster* and present both sets of statistics in Supplementary Excel Table 1.

Our assemblies were compared to the assembled reference sequence for each organism to determine % of genome covered, estimated duplication and number of mismatches and indels between them using QUAST version 5.0.2 (Gurevich et al., 2013).

Supplementary Table 1. The percentage of conserved genes identified in BUSCO analyses.

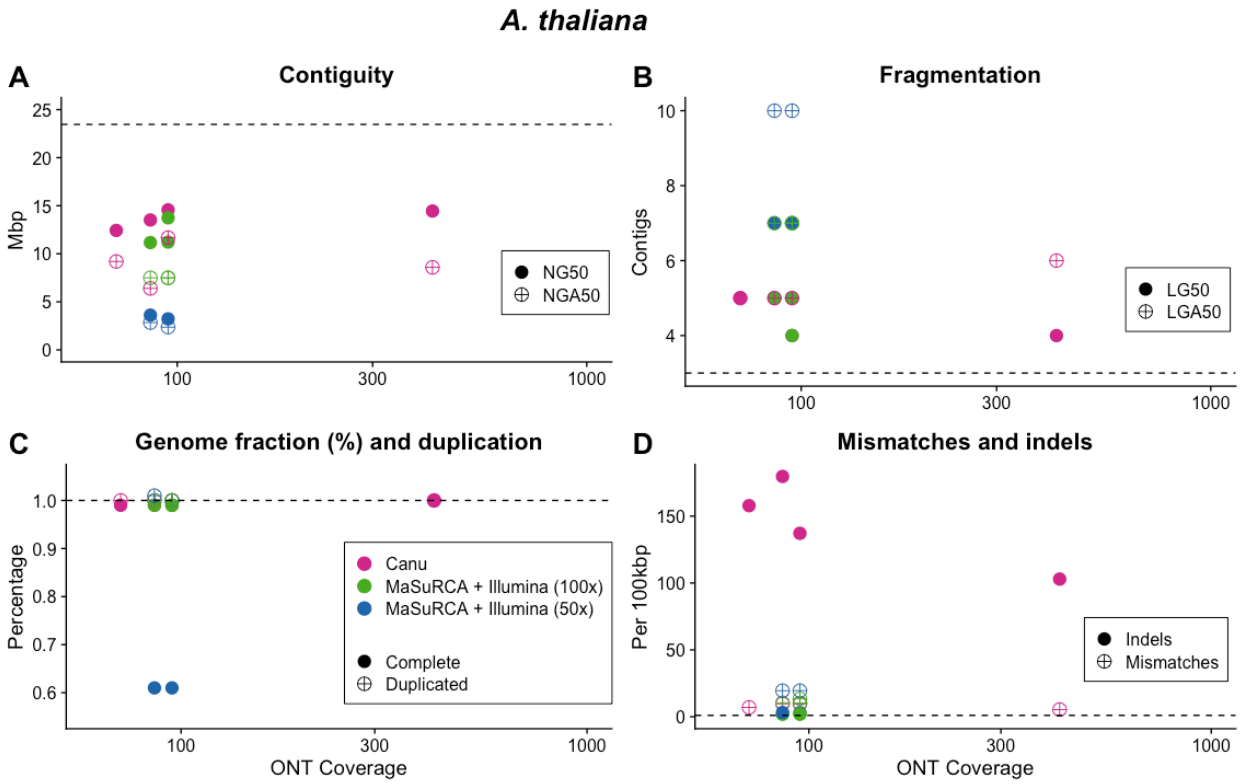
Sample	Strategy		Coverage		BUSCO			
			ONT	Illumina	Single (%)	Duplicated	Fragmented	Missing
A. thaliana		Canu	420	-	98.80	0.20	0.50	0.50
			86	-	98.40	0.50	0.70	0.40
			95	-	99.10	0.50	0.00	0.40
			71	-	98.60	0.50	0.50	0.40
		MaSuRCA	86	100	98.8	0.7	0	0.5
			95	100	99.1	0.5	0	0.4
			86	50	64	1.2	0.7	34.1
			95	50	64.5	0.7	0.7	34.1
			95	100	98.8	0.7	0	0.5
			-	100	98.1	0.9	0.5	0.5
	Pilon	Polish round 0	420	-	98.80	0.20	0.50	0.50
	corrected	1	420	-	99.1	0.5	0	0.4
	Canu	2	420	-	99.1	0.5	0	0.4
		3	420	-	99.1	0.5	0	0.4
		4	420	-	99.1	0.5	0	0.4
	Ensembl Reference		-	-	99.1	0.5	0	0.4
C. elegans		Canu	336	-	97.60	0.60	0.40	1.40
			264	-	96.80	0.50	0.60	2.10
			144	-	97.40	0.50	0.50	1.60
			72	-	97.60	0.60	0.30	1.50
			45	-	97.60	0.60	0.40	1.40
		MaSuRCA	336	200	97.7	0.6	0.4	1.3
			336	100	72.8	0.2	0.3	26.7
			264	200	22.7	0.5	0.2	76.6
			264	100	3	0	0.1	6.9
			144	200	97.8	0.6	0.4	1.2
			144	100	73.1	0.2	0.3	26.4

			72	200	97.8	0.5	0.4	1.3
			72	100	73.2	0.3	0.3	26.2
			-	200	97.60	0.60	0.40	1.40
	Pilon	Polish round 0	336	-	97.60	0.60	0.40	1.40
	corrected	1	336	-	97.7	0.6	0.4	1.3
	Canu	2	336	-	97.7	0.6	0.4	1.3
		3	336	-	97.7	0.6	0.4	1.3
		4	336	-	97.7	0.6	0.4	1.3
		Ensembl Reference	-	-	98	0.5	0.3	1.2
D. melanogaster		Canu	219	-	93.9	1.3	0.3	4.5
			108	-	91.30	0.90	2.20	5.60
			113	-	91.70	0.60	2.40	5.30
			60	-	86.60	1.30	4.10	8.00
		MaSuRCA	108	100	93.9	1.2	0.3	4.6
			113	100	94	1.2	0.3	4.5
			60	100	93.8	1.3	0.3	4.6
			51	100	93.9	1.2	0.3	4.6
			60	50	61.9	0.4	1	36.7
			51	50	61.8	0.4	1	36.8
			-	100	93.5	1.5	0.4	4.6
	Pilon	Polish round 0	108	-	91.30	0.90	2.20	5.60
	corrected	1	108	-	93.9	1	0.5	4.6
	Canu	2	108	-	94	1.2	0.3	4.5
		3	108	-	94	1.2	0.3	4.5
		4	108	-	94	1.2	0.3	4.5
		Ensembl Reference	-	-	94.1	1.2	0.3	4.4

Supplementary Table 2: Assembly Approaches for *Caenorhabditis remanei* PX356

Approach	ONT Coverage	Illumina Coverage	# of Contigs	BUSCO Completion (Single) %
Canu	102x	N/A	183	80.03/97.7*
MaSuRCA	102x	450x	366	96.6
AbySS†	N/A	450x	827	95.5

*After polishing with Pilon (Walker et al., 2014)
†From Fierst et al., 2015



Supplementary Figure 1. Canu assembled sequences had higher (A) contiguity; lower (B) fragmentation; greater (C) completeness and duplication and (D) a greater number of insertion/deletion errors when compared with MaSuRCA assembled sequences.