De novo assembly of prokaryotic and large eukaryotic genomes with long nanopore reads

The long-read capability of nanopore sequencing simplifies *de novo* assembly of microbial and eukaryotic whole genomes, resulting in increased assembly contiguity

Contact: **publications@nanoporetech.com** More information at: **www.nanoporetech.com** and **publications.nanoporetech.com**

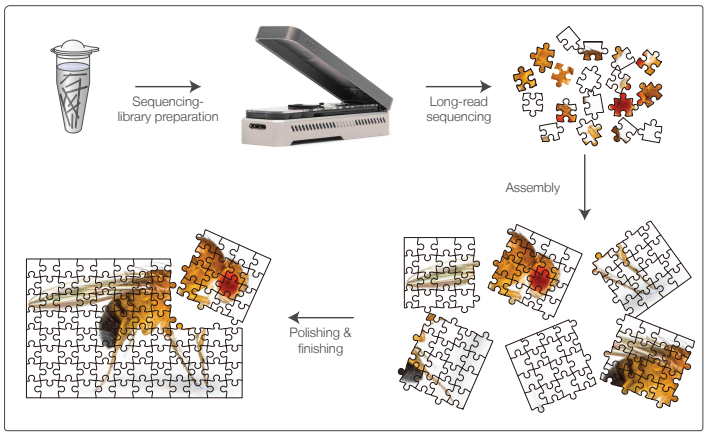

Fig. 1 Illustration of a typical approach to genome assembly

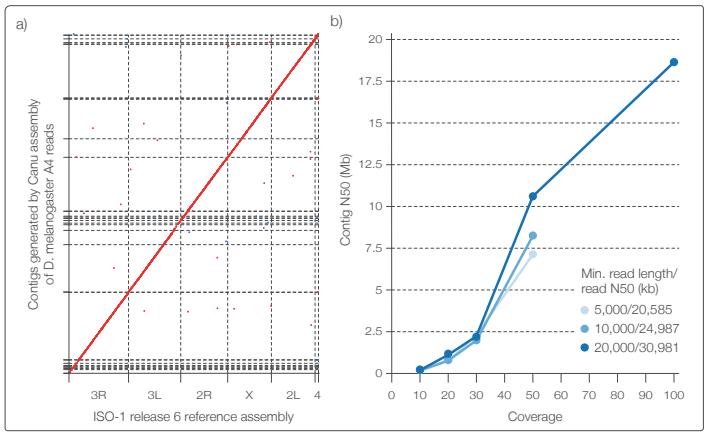

Fig. 2 Drosophila assembly a) MUMmer plot b) contiguity with different read N50s and coverage

Long nanopore reads enable *de novo* assembly of large and complex genomes

Nanopore reads can reach hundreds of kilobases in length, which is more than sufficient to span entire viral genomes in single reads. In contrast, to obtain a complete genome sequence from bacterial, or larger, genomes it is currently necessary to reconstruct the sequence by aligning and joining together overlapping sequence reads. This process is termed '*de novo* genome assembly' (Fig. 1). Assembling genomes using data from short-read sequencing technologies presents a computational challenge, and the results tend to be imperfect, particularly when the genomes contain extensive repetitive regions. Long reads make assembly far easier, and allow us to resolve repeats and structural variants that are several kilobases in length.

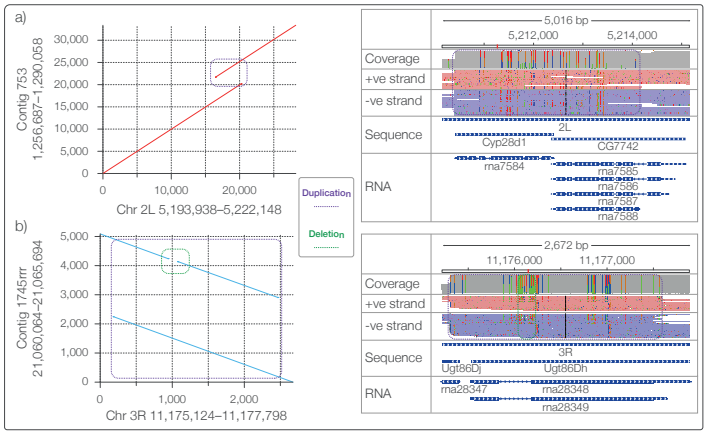

Fig. 3 Structural variation (SV) in the A4 assembly a) duplication b) duplication with deletion

Highly contiguous assembly of *Drosophila* strain A4 using long reads

We generated 40 Gb of sequence data from 200 female *D. melanogaster* A4 flies, and filtered for quality scores > 8. We randomly sampled this data to provide different levels of coverage and read N50s, and assembled with *Canu*. We polished with *Racon* followed by *Nanopolish*, using homopolymer fixing and methylation awareness. For comparison, we used the ISO-1 release 6 strain (Fig. 3a). The most contiguous assembly was from 100x of 1D reads > 20 kb, which gave an N50 of 18.3 Mb. The longest contig was ~ 23 Mb. Lower coverage corresponds to lower assembly contiguity, as would be expected. We observed a similar trend for lower read N50, and this effect is more pronounced when a higher read coverage is used in the assembly (Fig. 3b).

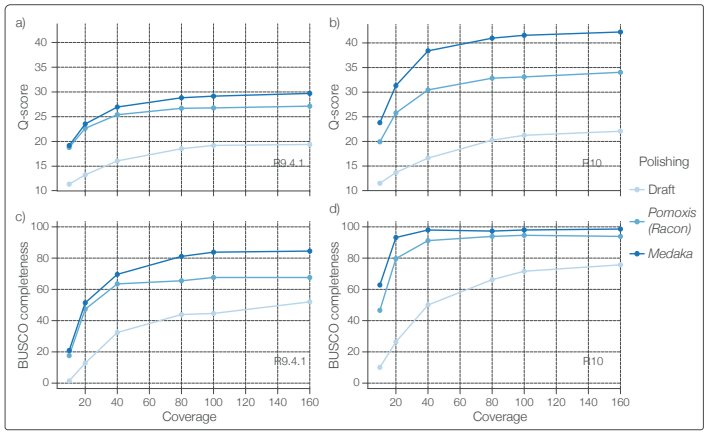

Fig. 4 Assembly of R10 data. *Pomoxis* and *Medaka* available from github.com/nanoporetech

Interpreting SV is essential for a complete understanding of genome architecture

De novo genome assemblies created from short read data tend to be fragmented, and can contain collapsed or misassembled repeat regions. The longer the sequence read, the longer the repetitive region or structural variant that can be resolved. Nanopore reads can be tens of kilobases in length, and can produce highly contiguous assemblies with correctly resolved structural variants. Fig. 4a shows a duplication of the *Cyp28d1* gene in the nanopore Drosophila A4 assembly, which is supported by an increase in read depth at the site of the gene. *Cyp28d1* is thought to be involved in the metabolism of insect hormones and the breakdown of synthetic insecticides. Fig. 4b shows a duplication of the mRNA 2834 on chromosome 3R, where one copy contains a deletion. Again, the duplication causes an increase in coverage in the IGV plot.

R10 pores show improved consensus accuracy and gene completeness

To investigate the effect of sequencing depth and polishing methods on assembly quality we analysed *E. coli* gDNA on R9.4 and R10 pores. Samples were sequenced to high depth and then subsamples were assembled using *Pomoxis*. Accuracy was assessed by comparing the alignment to the reference (Figs. 4a and 4b) and measuring the proportion of full-length single-copy genes predicted by BUSCO (Figs. 4c and d). NB genes present in the assembly in either partial or duplicated states are counted as 'missing' in this statistic. Metrics were calculated for the draft *miniasm* assembly, after *Racon* polishing and after ONT-specific error correction with *Medaka*. At 40x coverage the R10 dataset had the same completeness score as the reference genome (98.6%), while at 160x it achieved an accuracy of Q42.21 (99.994%).