Improved *de novo* assembly with nanopore ultra-long and duplex data, and scaffolding using Pore-C

Accurate, complete and contiguous genome assemblies are essential for identifying important structural and functional elements of genomes and for identifying genetic variation in an unbiased manner

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

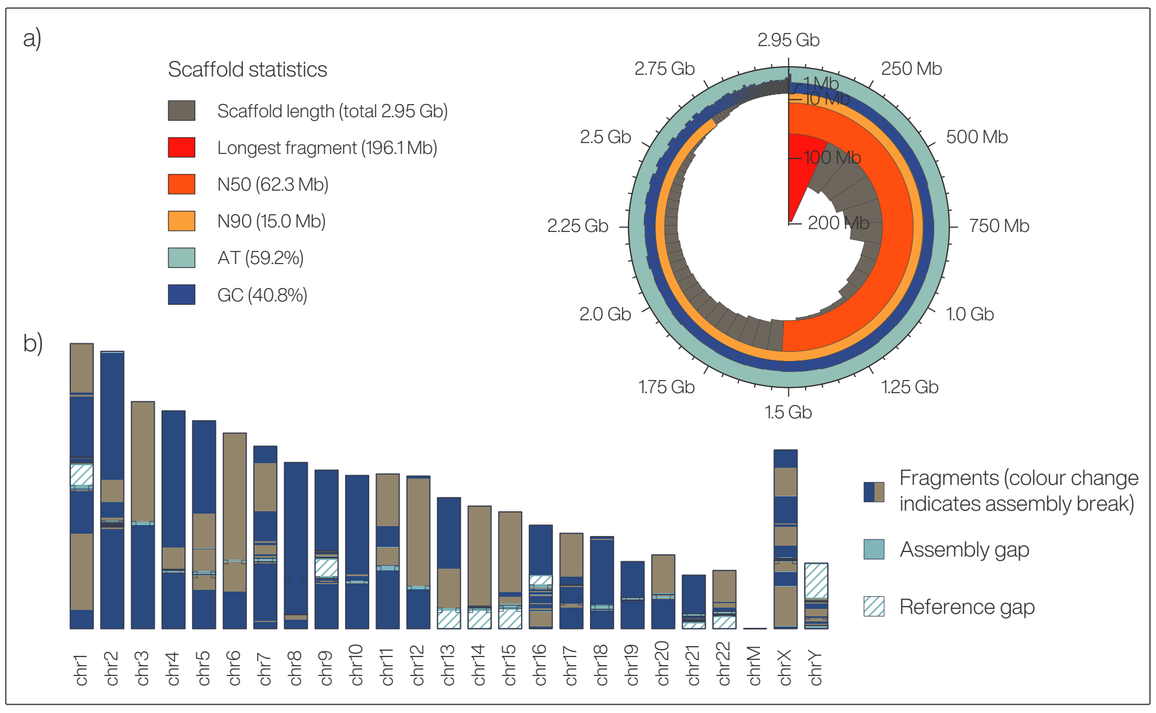

Fig. 1 Assembly of Hg002 a) assembly statistics b) contigs across the genome

Extremely contiguous assembly of human genome Hg002 using ultra-long reads

We used 60x of nanopore ultra-long reads (read N50 > 100 kb) base called with guppy 5 (SUP model) to produce a highly contiguous assembly of Hg002 using *Flye* 2.9. The final assembly has a contig N50 of 62 Mb (Fig. 1a). The largest contig was 196 Mb and 90% of the genome was contained in contigs > 15 Mb. The assembly showed high accuracy, yielding a BUSCO score of 96.9% (complete genes). Fig. 1b shows contig sizes along the chromosomes. Colour changes between beige and blue show contig or alignment breaks. For half the chromosomes, >95% of sequence assembled into 5 fragments or fewer. The efficiency of long-read assembly and polishing tools means that the overall runtime was <72 hours on a single AWS instance.

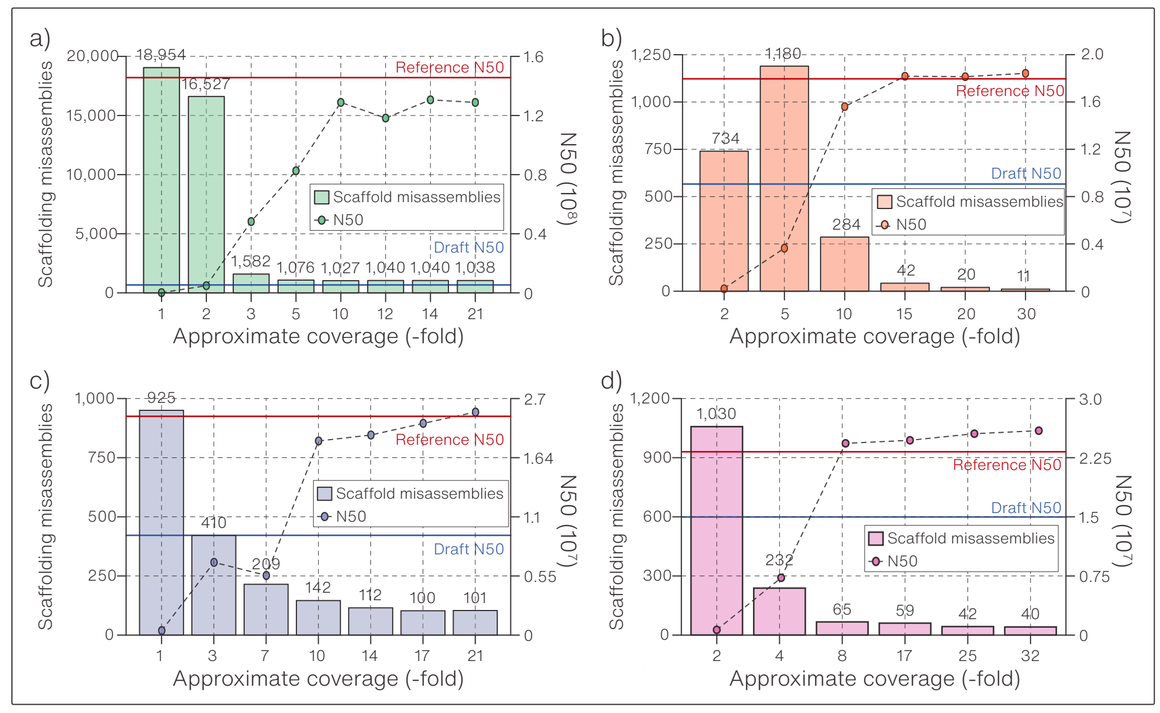

Fig. 2 Scaffolding genomes with Pore-C a) NA12878 b) C. elegans c) drosophila d) arabidopsis

Using Pore-C contact information to improve assembly contiguity of several genomes

To demonstrate the effectiveness of scaffolding assemblies with Pore-C data, we performed *de novo* assembly using *Flye* for all genomes apart from human (NA12878), which we assembled with *Shasta*. For each genome, Pore-C data was processed to create virtual paired contacts, which were used for scaffolding the assemblies. The results show that scaffolding with approximately 10x Pore-C data can increase assembly contiguity substantially, even when the initial draft assembly is highly fragmented (Figs. 2a-2d). Where scaffolded assembly N50 is greater than the reference, it indicates that sequence that is missing from the current reference assemblies.

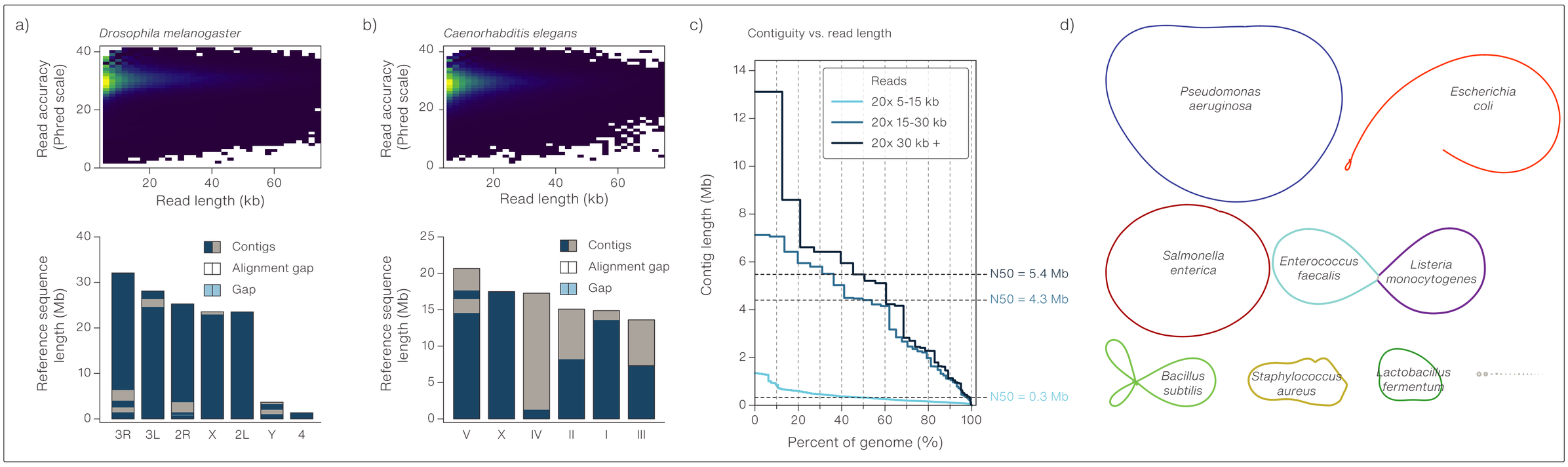

Fig. 3 Duplex assemblies of a) Drosophila melanogaster b) and c) Caenorhabditis elegans and d) the Zymo mock bacterial community

Getting the best of both worlds: long, high-accuracy duplex reads lead to improved assembly contiguity and accuracy across taxa

Oxford Nanopore's duplex pipeline finds reads originating from both strands of the same double-stranded DNA molecule by comparing reads which pass through the same pore in succession. Duplex base-calling combines signal information from both strands and can generate a 2-pass consensus read sequence with >99.9% modal accuracy (>30 on the Phred scale). Read accuracy from duplex base-calling is independent of read length (Figs. 3a and 3b, top), allowing for arbitrarily long highly accurate reads. These reads can be assembled with the latest generation of high-performance genome assembly tools, including *Hifiasm*, *HiCanu*, *Verkko* (Figs. 3a and 3b, bottom), and *La Jolla Assembler* (Fig. 3c), leading to near complete chromosome or chromosome-arm on single contigs (Figs. 3a and 3b, bottom). Longer read lengths unlock higher-contiguity assemblies even at low coverages, as can be seen in our assemblies of different length 20x subsets of C. elegans duplex data; assembly of 20x of duplex reads longer than 30 kb (read length N50 = 41 kb) led to an assembly with a 25% higher contig N50 compared with 20x of duplex reads between 15-30 kb (read length N50 = 22 kb) and an order of magnitude higher contig N50 compared with 20x of duplex reads between 5-15 kb (read length N50 = 10 kb). High accuracy reads also significantly improve assembly of samples for which obtaining high molecular weight DNA continues to be challenging, as in some bacteria. For example, a 50-200x duplex dataset for the Zymo mock community with read N50s of ca. 5 kb was mostly assembled to single, nearly perfect circular contigs by the *MetaFlye* assembler (Fig. 3d).