DNASTAR
SOFTWARE FOR LIFE SCIENTISTS

LASERGENE

De Novo Long-Read Assembly Benchmarks Using SeqMan NGen 19.0

Recent advances in long-read sequencing have dramatically expanded the scope of *de novo* assembly, making it practical to assemble not only microbial genomes, but also large eukaryotic genomes. Lasergene Genomics leverages these advances with an enhanced long-read *de novo* assembly workflow in SeqMan NGen. Optimized for PacBio HiFi and Oxford Nanopore sequencing data, SeqMan NGen delivers highly contiguous assemblies with minimal user intervention.

Our benchmark testing demonstrates excellent assembly performance across a wide range of genome sizes. Contig N50 values are the length cutoff for the longest contigs that contain 50% of the total genome length. Compared to *de novo* assemblies generated from short-read data, which often produce Contig N50 values below 100 Kbp, long-read assemblies generated with SeqMan NGen typically exceed 1 Mbp.

Benchmark Testing Setup

Software:

- SeqMan NGen v19.0 *de novo* long-read workflow using default parameters.

Hardware (including minimum recommendations):

- 2025 Mac Studio
- 28-core Apple M3 Ultra (we recommend at least 8 cores)
- 96 GB RAM (≥32GB RAM is needed for data sets with >50Gbp input reads)
- 4TB high-speed NVMe SSD (we recommend using an external data drive)

Data:

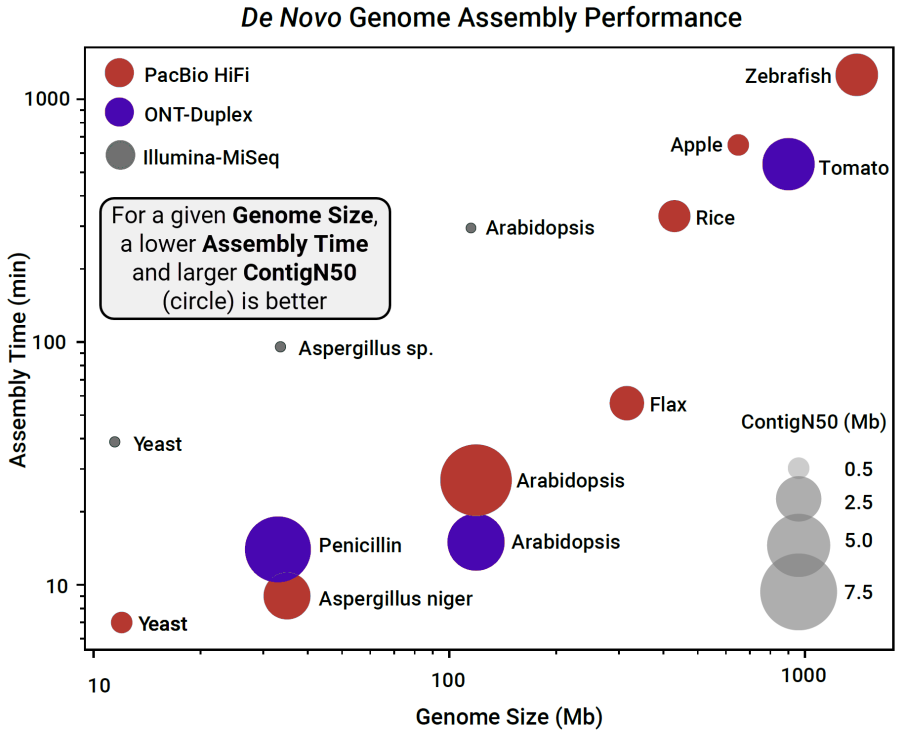
- Publicly available genomic data from eukaryotic species ranging in size from 30Mbp (*Aspergillus*) to 1.4 Gbp (zebrafish)
- Separate assemblies were run using PacBio HiFi, Oxford Nanopore Duplex, and Illumina MiSeq (301bpx2) data

Benchmark Testing Results

De Novo Long-Read Benchmarks

The graph at right shows the linear relationship between log(Genome Length) and log(AssemblyTime) for all three data types tested: PacBio HiFi, ONT Duplex and Illumina MiSeq.

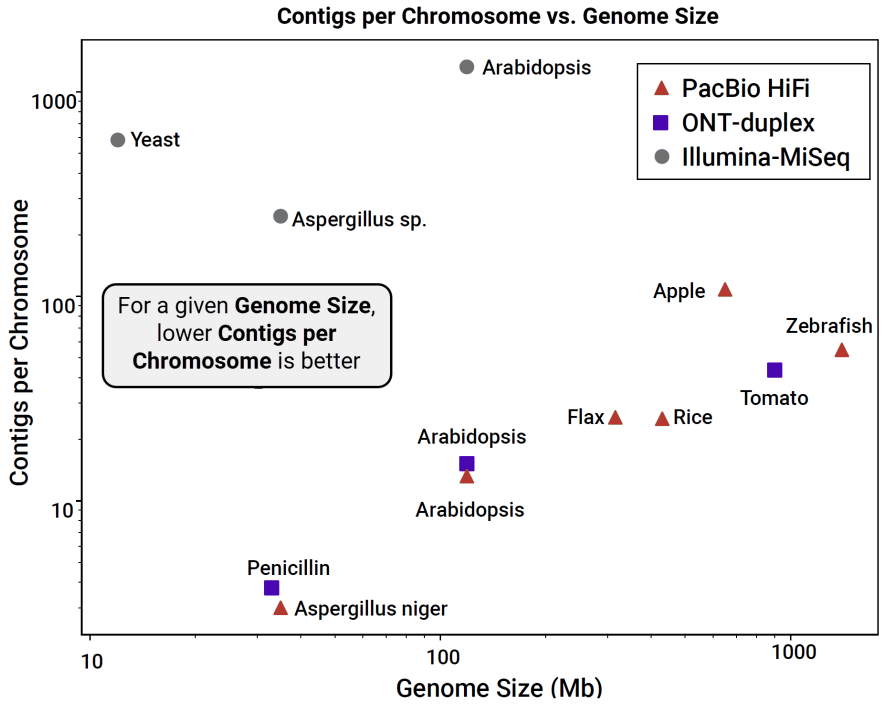
"Bubbles" represent ContigN50 values. Larger bubbles are better, as they indicate more contiguous (less fragmented) assemblies. For a given genome size, long-read assemblies were faster (lower on the graph) and produced fewer and much larger contigs (larger bubbles) than Illumina data. Eukaryotic genomes contain many repetitive elements that are difficult to resolve with short reads.



Contigs Per Chromosome vs. Genome Size

The graph at right shows the number of contigs per chromosome plotted against genome size, with lower contig numbers (data points closer to the bottom of the graph) being better. Because eukaryotic chromosomes contain a large centromere that cannot be sequenced, two contigs per chromosome is the best possible result.

Results for *Arabidopsis* and *Aspergillus* highlight the orders-of-magnitude difference in contig numbers for *de novo* assemblies built with Illumina MiSeq data versus long-read data.



608.258.7420 USA
866.511.5090 toll free
0.808.271.1041 UK
0.800.182.4747 Germany
1202 Ann Street
Madison, WI 53713
www.dnastar.com
info@dnastar.com