DNASTAR®
Software for Life Scientists

Genome Polishing Benchmarks: DNASTAR's SeqMan NGen vs. three open-source tools

In genome "polishing," assembly software searches for local misassemblies and other inconsistencies in a draft genome assembly and then corrects them. This paper compares accuracy and other statistical benchmarks for SeqMan NGen's "short read polishing of a long-read draft genome" workflow versus several open-source tools.

The first step in the genome polishing workflow was to use Canu to create draft genome assemblies for six eukaryotic and prokaryotic species from long-read sequencing data. Illumina reads were then utilized to polish the assembly using one of three tools: SeqMan NGen, Pilon or SPAdes. The unpolished Canu assemblies were also included in the comparison.

Our results demonstrate that SeqMan NGen beat or tied the alternative tools in 18 of 24 statistical metrics, while the next-best tool did so in only 11 of 24 metrics. SeqMan NGen produced the most accurate consensuses, consistently captured the highest percentage of the genome, and maximized the aligned length. In addition, SeqMan NGen was the fastest to install and use, and was the only tool that produced a fully-editable assembly.

Tools

All work was performed on Macintosh computers (40 GB or 48 GB RAM) running macOS Catalina 10.15 or macOS High Sierra 10.13.

The assembly tools used were:

- SeqMan NGen 17.0.2.2 – Part of Lasergene Genomics, DNASTAR, Inc.
- Canu 1.9 - Celera and Pacific Biosciences
- Pilon 1.23 - Broad Institute
- SPAdes 3.14.0 - Center for Algorithmic Biotechnology

Input Data

Six different data sets were selected:

- *Escherichia coli* str. K-12 substr. MG1655
- *Fusobacterium nucleatum* subsp. *nucleatum* str. ATCC 25586
- *Fusobacterium periodonticum* str. 2_1_31
- *Klebsiella pneumoniae* str. INF042
- *Pseudomonas koreensis* str. P19E3
- *Saccharomyces cerevisiae* str. S288c

Three types of data—each originating from the same strain—were required for each set.

- Long reads for Canu assembly were all Oxford Nanopore Technologies (MinION) .fastq reads except for the *P. koreensis* data, which consisted of Pacific Biosciences (PacBio) .fastq reads. Except for the *E. coli* MinION data, which was from Nick Loman's lab at the University of Birmingham, all read data was retrieved from NCBI's Sequence Read Archive (SRA).
- Illumina reads for polishing came from NCBI's Sequence Read Archive (SRA).
- Complete reference assemblies for comparison and assessment were GenBank files downloaded from NCBI (or RefSeq files, in the case of *S. cerevisiae*). For each multiple-replicon genome, individual files were combined into a single multi-sequence file. DNASTAR's SeqNinja application was used to generate .fasta files from the GenBank files.

Analysis Workflow

The analysis workflow involved the following steps for each data set:

1) Canu was used to create draft assemblies of the MinION or PacBio data using default parameters.
2) For *S. cerevisiae* and *P. koreensis* only, the Canu contigs were reordered using Mauve 2.4.0 (Darling Lab, University of Technology Sydney) to simplify downstream analysis. No corrections were made for circular permutations, etc.

3) Contigs from the draft Canu assemblies were polished with Illumina data using one of three different tools:

   - Pilon uses one or more BAM files of reads aligned to the draft contigs. This required mapping the Illumina reads to the Canu contigs using Bowtie2 2.3.5.1 (open source), converting the SAM files into sorted BAM files using Samtools 1.10 (open source) and then running Pilon with default parameters.

   - SPAdes can do hybrid assemblies. For comparison purposes to the other polishing workflows, however, the Canu contigs were first set to the "trusted contigs" option and used for graph construction, gap closure and repeat resolution. Next, the "careful" option was used to try to reduce the number of mismatches and short indels.

   - The SeqMan NGen workflow "Short read polishing of a long read draft genome" was used with default parameters.

4) Any non-ACGTN characters were converted to Ns to avoid errors.

5) The contigs from each of the polishing protocols were compared to the corresponding complete genomes. This was done using QUAST 5.0.2, a quality assessment tool for evaluating and comparing genome assemblies.

Calculations

Genome statistics are shown in Table 1 and were calculated as follows:

- **Genome fraction (%)** = the percentage of aligned bases in the reference genome; used as a measure of how well the reference is covered. Contigs from repetitive regions may map to multiple places and may be counted multiple times.

- **Total aligned length** = the total number of aligned bases in the assembly; a measure of how well the assembly corresponds to the reference. Usually smaller than the total assembly length because some of the contigs may be unaligned or partially unaligned.

- **Number of contigs** = The final number of contigs produced during the assembly workflow. Fewer contigs is preferable for downstream analysis and genome closure.

- **Accuracy** = $100 * (1 - \frac{\text{combined length of mismatches and indels}}{\text{total aligned length}})$

Table 1. Genome statistics for four assembly tools and six genomes

	E. coli				*F. nucleatum*			
Genome statistics	**Canu**	**Pilon**	**SPAdes**	**SeqMan NGen**	**Canu**	**Pilon**	**SPAdes**	**SeqMan NGen**
Genome fraction (%)	99.9	99.9	99.1	**100**	100	100	97.9	**100**
Total aligned length	4,607,087	4,632,861	4,595,257	**4,638,000**	2,193,884	2,216,938	2,134,351	**2,229,176**
No. of contigs	1	1	743	**5**	1	1	1802	**31**
Accuracy	99.4	100	100	**100**	98.9	99.9	100	**100**

	F. periodonticum				*K. pneumoniae*			
Genome statistics	**Canu**	**Pilon**	**SPAdes**	**SeqMan NGen**	**Canu**	**Pilon**	**SPAdes**	**SeqMan NGen**
Genome fraction (%)	99.5	99.5	97.3	**99.9**	100	100	99.2	**100**
Total aligned length	2,503,481	2,530,305	2,470,727	**2,540,942**	5,540,312	5,590,941	5,416,830	**5,593,498**
No. of contigs	34	34	442	**5**	3	3	144	**6**
Accuracy	98.8	99.9	99.9	**100**	99.0	99.9	100	**100**

	P. koreensis				*S. cerevisiae*			
Genome statistics	**Canu**	**Pilon**	**SPAdes**	**SeqMan NGen**	**Canu**	**Pilon**	**SPAdes**	**SeqMan NGen**
Genome fraction (%)	95.7	95.7	96. 7	**98.5**	99.3	99.3	96.2	**99.3**
Total aligned length	7,244,739	7,246,200	7,249,865	**7,472,809**	12,420,135	12,604,157	12,021,367	12,590,874
No. of contigs	12	12	75	**70**	75	75	3874	**129**
Accuracy	100	100	100	**100**	98.2	99.7	99.9	**99.9**

* Green shading denotes cases where SeqMan NGen produced the best result for a given polishing metric, or where there was a tie for best result between SeqMan NGen and another tool.

Discussion

Genome fraction and **Total aligned length**: SeqMan NGen consistently outperformed all three of the other tools in capturing the highest genome fraction. The average **genome fraction** for each tool was 99.6% (SeqMan NGen), 99.1% (Pilon and Canu) and 97.7% (SPAdes). The numerical difference may seem small, but the difference between the highest and lowest scoring tools can be seen clearly in the **Total aligned length** statistic, where SPAdes failed to find nearly 223,000 base pairs of the genome compared to SeqMan NGen.

Number of contigs: While fewer contigs are generally preferred for downstream analysis and genome closure, that can also be an over-simplification. If there are false joins, for example, contigs are harder to take apart than to merge during manual finishing. Canu and Pilon created the fewest contigs on average. Note that Pilon doesn't add contigs, but also doesn't merge or delete them. In attempting to improve the assembly, Spades tended to break up existing contigs, while SeqMan NGen tended to merge them.

Accuracy: SeqMan NGen had the highest average accuracy (over 99.98%) and achieved an accuracy ≥ 99.9% for all six data sets. Pilon had the next best average accuracy (99.9%), followed by Canu (99.1%) and SPAdes (97.7%).

Conclusion

While long read data is essential for overall coverage and for dealing with repetitive regions that confound assembly, their accuracy leaves much to be desired. Whether done during preassembly or after an initial *de novo* assembly, some cleanup is required to avoid errors such as small scale misassemblies that can lead to base errors and small indels.

Our benchmarking data showed that SeqMan NGen's genome finishing workflow tied or beat the other three tools in 18 of 24 statistical metrics (4 statistics for each of 6 species). SeqMan NGen had the highest accuracy, assembled a larger percent of the genome and created many fewer contigs than SPAdes. By comparison, Pilon won or tied in 11 metrics; Canu in 9; and SPAdes in 5.

In addition, SeqMan NGen required the fewest steps to install and its wizard interface also made it the fastest genome polisher to run. By comparison, all three open-source tools required the user to set up a particular working environment or circumvent an operating system's security provisions during installation or use.

Finally, the three open-source tools produce just the consensus .fasta sequences. The user cannot inspect or adjust the alignments or add additional data. SeqMan NGen is the only one of the four tools evaluated to output an editable assembly file. Its .sqd output file can be further edited and polished manually using DNASTAR's SeqMan Pro and/or SeqMan Ultra applications.