Microbial Data Analysis for Uncorrected Long Reads with QIAGEN® CLC Genomics Workbench

Rune M. Friborg[1], Jonathan Jacobs[2], Leif Schauser[1]

[1]QIAGEN Digital Insights, Silkeborgvej 2, 8000 Aarhus C, Denmark; [2]QIAGEN Digital Insights, 1001 Marshall Street, Redwood City, CA 94063, USA

Abstract

The QIAGEN CLC Genomics Workbench now enables analysis of uncorrected long reads from Oxford Nanopore and PacBio® using a newly developed plugin that provides seamless integration with other QIAGEN CLC products.

The Long Read Support plugin builds on established state-of-the-art pipelines for read mapping, read error correction, de novo assembly and contig polishing, that are made available through the user-friendly workbench Graphical User Interface (GUI), with no command-line or code compilation needed.

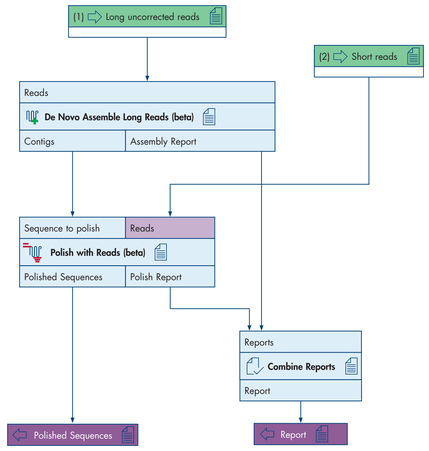
Benefits of the Long Read Support plugin include:

- Combination of best practices for polishing and assembly of long reads (1, 2)
- Easy to use GUI interface for non-experts
- Execute on a laptop
- 100% cloud ready
- Prebuilt or user-designed custom workflows
- Inbuilt data security and data compression
- Full support for enterprise integration

Building Custom Workflows for Hybrid Assembly (1)

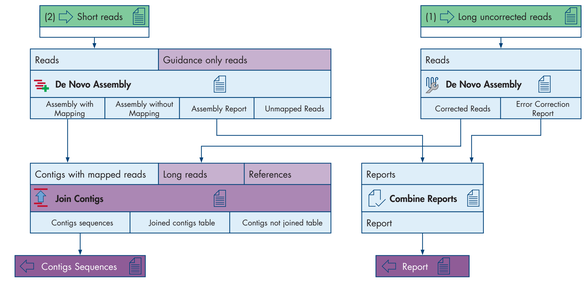
It is now possible to use long reads or short reads as the primary basis for assembly.

Scaling NGS analysis requires the ability to build complex version-controlled workflows. In addition, workflows must be easily sharable, enable third-party applications, and flexibly execute in situ or in the cloud from the GUI or the command-line. The CLC platform supports all of these requirements.

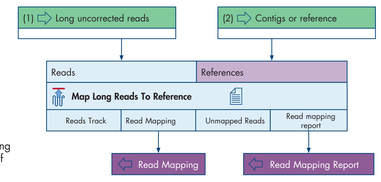


A workflow for de novo assembly of uncorrected long reads, followed by polishing with short reads.

Building Custom Workflows for Hybrid Assembly (2)



An alternative workflow for hybrid assembly. First, de novo assembly of short reads, followed by joining contigs using corrected long reads output from the new Correct Long Reads tool.



Map long reads to references. Map against large genomes, or, after assembling long reads to contigs, map back reads to contigs to verify even coverage of assembled genome(s).

Access to Minimap2, Miniasm and Racon in One Easy-to-Use Plugin

Overlap layout consensus (OLC)-based de novo assembly of noisy long reads.

Minimap2 (3)

- Aligner exceptionally good at mapping long reads from Oxford Nanopore or PacBio instruments.
- Can be used to align reads against reads and produce an overlap alignment, as well as map reads against contigs or references.

Miniasm (4)

- Fast OLC-based de novo assembler for uncorrected long reads.
- At an early development stage, but shows stability and value when benchmarked against other long read assembly pipelines.

Racon (5)

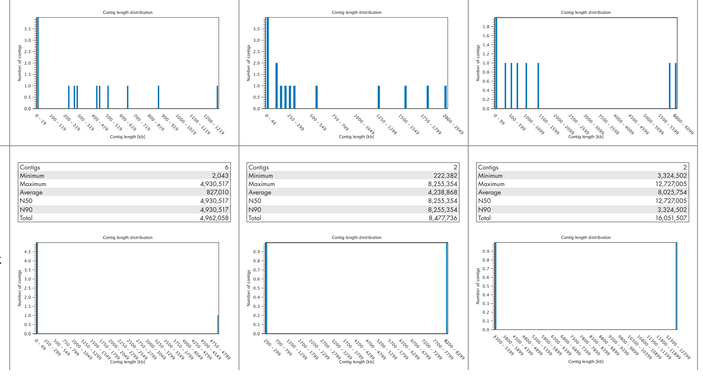
- Powerful consensus module that can be used to polish or correct sequences using either short reads or noisy long reads.
- Works particularly well combined with Miniasm.

The "De Novo Assemble Long Reads" tool executes the pipeline:

1. All reads are mapped against each other using Minimap2.
2. Miniasm loads the read mapping from Minimap2 and produces raw contigs based on overlaps and coverage.
3. Racon polishes the raw contigs using the provided reads and a read mapping where the reads are mapped against the raw contigs.
4. Racon is run a second time to polish the contigs from the previous step, as more reads are mapped successfully to the polished contigs, thus increasing quality of the final contigs.

Assemble the Complete Genome With Long Reads

	Salmonella enterica	*Bradyrhizobium* spp. BTAi1	*Minicystis rosea*
50x	Contigs 13 Minimum 1,186 Maximum 1,288,344 Average 383,122 N50 656,861 N90 277,579 Total 4,980,460	Contigs 15 Minimum 9,177 Maximum 2,017,743 Average 567,507 N50 1,598,862 N90 287,825 Total 8,527,603	Contigs 9 Minimum 1,017 Maximum 6,027,044 Average 1,784,133 N50 5,898,537 N90 765,876 Total 16,057,200
150x	Contigs 6 Minimum 2,043 Maximum 4,930,517 Average 827,010 N50 4,930,517 N90 4,930,517 Total 4,962,058	Contigs 2 Minimum 222,382 Maximum 8,255,354 Average 4,238,868 N50 8,255,354 N90 8,255,354 Total 8,477,736	Contigs 2 Minimum 3,324,502 Maximum 12,727,005 Average 8,025,754 N50 12,727,005 N90 3,324,502 Total 16,051,507



Conclusions

The new Long Read Support plugin to the CLC Genomics Workbench provides convenient analysis of long reads on the CLC-powered platform:

- Easy access to public and private reference data
- Standardized, customizable and sharable workflows
- A host of NGS tools
- Enterprise level scaling on any computational platform using GUI

References

1. Wick, R.R., and Holt, K.E. (2019) Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 1; peer review: 1 approved]. F1000Research **8**, 2138. (https://doi.org/10.12688/f1000research.21782.1)
2. Senol Cali, D., et al. (2019) Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. Brief Bioinform. **20**, 1542.
3. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**, 3094.
4. Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics **32**, 2103. https://doi.org/10.1093/bioinformatics/btw152
5. Vaser, R., et al. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. **27**, 737.

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at **www.qiagen.com** or can be requested from QIAGEN Technical Services or your local distributor.

1120195 01/2020

Sample to Insight