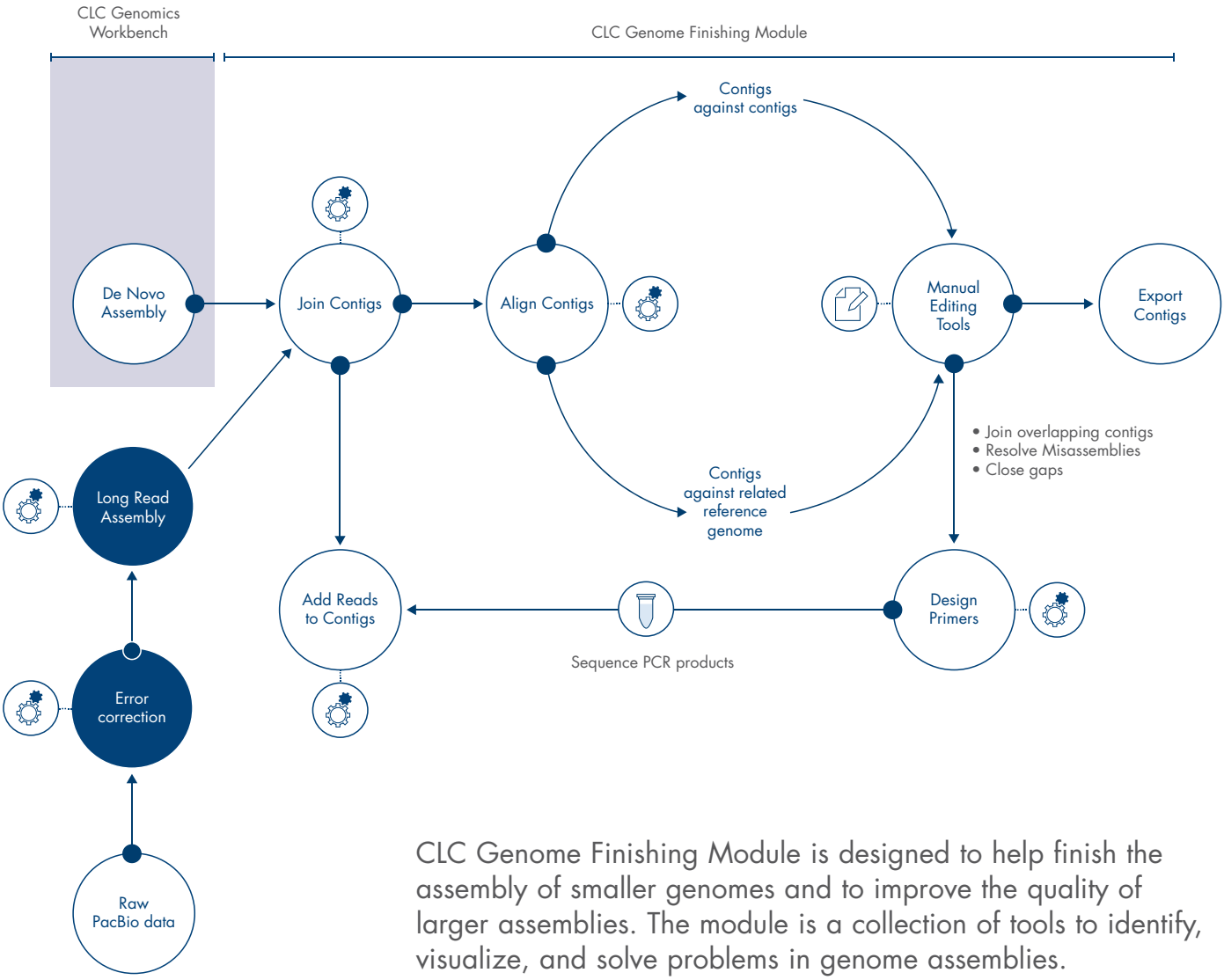
CLC Genome Finishing Module

Included features and tools



CLC Genome Finishing Module is designed to help finish the assembly of smaller genomes and to improve the quality of larger assemblies. The module is a collection of tools to identify, visualize, and solve problems in genome assemblies.

Genome finishing

High-throughput sequencing technologies have enabled rapid sequencing of whole genomes. However, short read lengths and repetitive sequences often result in fragmented assemblies and complicate genome finishing. CLC Genome Finishing Module is an add-on to CLC Genomics Workbench, designed to facilitate the often daunting and work-intensive multi-step process of genome finishing.

Get to high quality reference genomes in fewer steps

Automation where possible
The module automates steps like scaffolding, contig joining, and the ordering of contigs and scaffolds relative to each other or a closely related reference genome.

Manual editing where necessary
The above steps improve the outcome of the initial assembly. Remaining unresolved regions can further be investigated with the help of tools like the Analyze Contigs tool, enabling the user to visually inspect and improve the results.

Back to the bench
To resolve repetitive regions it can be necessary to amplify and resequence the respective genomic regions. CLC Genome Finishing Module makes it easy to design useful primer combinations and to improve the assembly once new data are available.

Scaling up your analysis. CLC Genome Finishing Module Server Extension
This extension to CLC Genomics Server allows users to run the tools offered by CLC Genome Finishing Module on a high performance computer or compute cluster.
Users enjoy the familiar intuitive user interface of CLC Genomics Workbench, while taking full advantage of the scalability of CLC Genomics Server.

Which assemblie sizes can be improved?
CLC Genome Finishing Module was designed for finishing the assemblies of smaller genomes, and is ideal for assembling microbes, eukaryotic parasites, or even fungi. The automated tools for scaffolding or contig joining can also improve the results of large genome assemblies, however manual editing is not feasible for large plant or animal genomes.

Tools included in CLC Genome Finishing Module

These tools can be combined in different ways.

Automated improvement of assembly quality:

The **Join Contigs** tool is designed to improve assembly quality where possible, and to reduce the number of contigs. This tool carries

Supported Assembly types

The module supports the common assembly types:

- Short read assemblies
- Hybrid assemblies combining short and long read data (e.g. Illumina, 454, and PacBio data)
- Pac Bio long read assembly. Raw PacBio reads are error corrected and assembled fast and compute resource-efficiently.

out automatic contig joining and scaffolding, leveraging paired read data or long reads. Optionally contigs can be aligned to each other or to closely related reference genomes.

The **Align Contigs** tool is the gateway to downstream visual inspection and manual editing of the contigs. An alignment of contigs is performed using BLAST, either against a reference sequence, or, if no reference sequence is available, the contigs themselves. This helps in determining the orientation and location of the contigs which allows the identification of possible misassemblies, repeats, and overlaps between contigs.

The tools **Correct PacBio Reads (beta)** and **De Novo Assemble PacBio Reads (beta)** error correct and assemble raw PacBio data into high quality assemblies. For microbial genomes the included preconfigured workflow **PacBio De Novo Assembly Pipeline** often directly yields gold-standard assemblies.

Tools for manual editing:

Identification and annotation of problematic regions in the de novo assembly is done with the **Analyze Contigs** tool. Regions are annotated when they have low or high coverage or sudden changes in coverage as well as single stranded or nonspecific regions, broken pairs or unaligned ends. Subsequently, these annotations can be used to pinpoint misassembled regions during visual inspection of the alignments.

A range of actions can be taken after selecting any area of the contig. The Analyze Contigs tool will mostly be used in combination with the Align Contigs tool.

The **Collect Paired Read Statistics** tool identifies paired reads between contigs, which in turn can help to determine the order and orientation of contigs. The tool provides information about the orientation of one contig relative to its mate contig, as well as the size of potential overlaps and unknown gaps between contig pairs.

The **Extend Contigs** tool helps to extend contigs with reads that continue beyond the ends of the contigs. The result of extending the contigs is that large overlaps are created between contigs. Using the Align Contigs tool on the extended contigs can then help to visualize overlapping contigs that can be joined.

The **Reassemble Regions** tool adjusts the read mapping and makes changes in the consensus sequence based on the reads in the selected region only. Hence, although the tool is not always capable of fixing problems in the assembly, the Reassemble Regions tool can

be used as an alternative to manual editing of sequences when problematic regions are encountered in the contigs.

Resequencing of unresolved regions:

The **Create Amplicon** tool can subdivide a problematic region, for example a region without any or with low coverage, into amplicons of suitable sizes and annotate these accordingly. Subsequently, these annotations can be used as targets for the Create Primers tool. The resulting primers are presented in a tabulated format.

The **Create Primers** tool is convenient whenever further sequencing is required, e.g. resequencing of regions with poor read quality, repeats, or low coverage. It can also be used to design edge primers for all input sequences.

Additional reads are added to an existing contig with the **Add Reads to Contigs** tool. The advantage of adding reads to existing read mappings, rather than making a new read mapping of both old and new reads, is that modifications that have already been made are preserved. This is particularly relevant after resequencing of problematic regions.

Other useful tools:

- Find Sequence
- Annotate from reference
- And more

For a trial visit:
qiagenbioinformatics.com/products/clc-genome-finishing-module/