



CLC Assembly Cell

Features & Benefits

All Major Sequencing Platforms Supported

- Illumina's Genome Analyzer
- SOLiD by Life Technologies
- 454 GS flx by Roche
- Helicoscope by Helicos
- Sanger sequencing data

System Requirements

- Mac OS X 10.4 or later (incl. Intel-based Macs)
- Windows 2000, Windows XP, Windows Vista or Windows 7
- Linux: Redhat or SuSE
- 32 bit version and 64 bit version of operating system/computer on all platforms
- 2 GB RAM required
- 16-48 GB required for large assemblies

Version 3.1 for Windows, Mac OS X, and Linux
CLC bio©Copyright 2010

Helping overcome the challenges of Next Generation Sequencing

CLC Assembly Cell is a command-line program that makes your high-throughput sequencing data analysis pipeline fast, flexible and easy to maintain.

Next Generation Sequencing technologies present a number of challenges to bioinformatics in terms of data analysis. The first challenge is assembling the data. Existing software for assembly typically supports only one type of data (either short or long reads) which makes it hard to exploit the advantages of combining different platforms. In contrast, CLC Assembly Cell offers high-speed assembly of sequencing data from all platforms (see text to the left). The ability to combine these technologies in the same analysis is one of the major strengths of the CLC Assembly Cell, as well as the ability to analyze reads ranging from 35bp to many hundred bp.

The Assembly Cell supports reference assembly and *de novo* assembly of genomes of all sizes.

De novo assembly

The *de novo* assembly of CLC Assembly Cell supports both short read and long read assembly, including 454/Titanium. CLC Assembly Cell also supports *de novo* assembly of paired end data (see figure 1 below) and hybrid assembly of multiple data types.

Reference assembly

CLC Assembly Cell supports reference assembly of Illumina Genome Analyzer, SOLiD, 454, Helicos and Sanger sequencing data.

It also handles short and long read assembly (incl. 454/Titanium), as well as gapped and ungapped alignment.

Human genome *de novo* assembly benchmarks

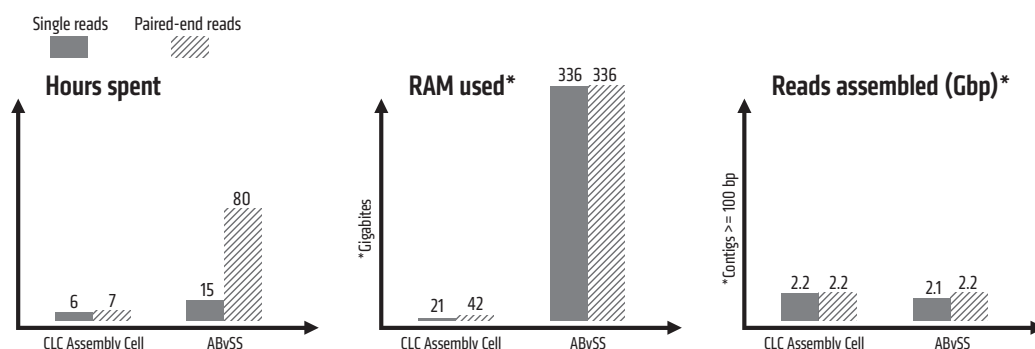


Figure 1: The data set used for the human genome assembly is produced with Illumina's Genome Analyzer and consists of approximately 3.6 billion reads resulting in 38 fold coverage of the human genome. Most reads have a length of 36 bp, summing up to 130 Gbp in total.

High performance

Due to utilization of SIMD instructions to parallelize and accelerate the analysis, CLC bio's *de novo* assembly algorithm is by far the fastest assembler at present. On large assemblies, e.g. *de novo* assembly of human genomes or plants, using paired-end data, the CPU-time is most often less than 1/10 of other assemblers.

The dramatic increase in speed does not mean compromising on memory consumption or quality of the assembly (the CLC Assembly Cell assembles more reads).

A special feature of the CLC Assembly Cell is that it automatically adjusts to the amount of memory available, so even large assemblies can be performed on computers with few GB of RAM. For example you can assemble a human genome on a single computer with 8 processor cores and 48 gigabytes of RAM in only 7 hours.

However, the assembly will run faster if there is enough memory available (see system requirements).

Reporting and statistics

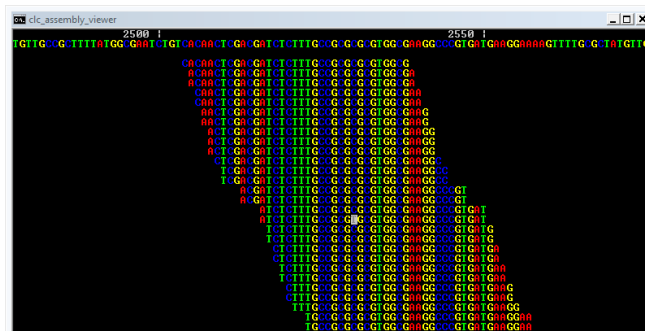
Besides the actual assembly, the CLC Assembly Cell includes a suite of assisting programs to help you manage and analyze assemblies. You can get detailed information about coverage, paired-ends matching and statistics on matched and non-matched reads.

The suite of assisting programs also includes a basic viewer to look at the alignment, and programs to divide and join assemblies. This makes it possible to do detailed analysis of smaller parts of the assembly and also to merge two assembly runs into one.

Paired-ends and mixed data sets

Data sequenced using paired-ends protocols provide much more information than single reads. The CLC Assembly Cell supports assembly of paired-ends data of all sequencing platforms. There is complete flexibility to specify the

distance between each read in the pair, and you can also perform assemblies of data sets that are sequenced using different paired-ends protocols as well as single reads data sets.



The command-line interface of CLC Assembly Cell enables the functionalities to be easily included in scripts and other Next Generation Sequencing work-flows. For example here is a "graphical" ASCII art assembly viewer to get a quick overview.

There are detailed statistics on the matching of pairs (how many reads were matched as pairs and non-pairs, why only one read in the pair match, etc.). In addition, you can see statistics on the distance between the paired reads.

Cluster support

CLC Assembly Cell supports and utilizes multi-core computers. For additionally parallel assembly on multiple computers in a cluster, CLC bio provides customized solutions for distributing the reference assembly calculations. Due to the flexibility of joining assembly results, it is very easy to distribute the assembly to multiple computers.

Visualization & downstream analysis

In order to support smooth workflows, CLC Assembly Cell is fully integrated with CLC Genomics Workbench, which delivers a large number of downstream analyses and can be used for advanced visualization of the assemblies produced with CLC Assembly Cell.

Contact your local sales representative or send an e-mail to sales@clcbio.com if you would like to try a free trial of CLC Assembly Cell.

CLC bio · Finlandsgade 10-12
Katrinebjerg · DK-8200 Aarhus N
Denmark
Phone: +45 70225509
www.clcbio.com · info@clcbio.com

CLC bio LLC · 245 First Street · Suite 1826
Cambridge · MA 02142
USA
Phone: +1 (617) 444-8765
www.clcbio.com · info@clcbio.com

MORE INFORMATION



www.clcbio.com/barcode

