Technical Note

De novo assembly

Sequencing technologies are in continuous development through emerging new technologies, improvements in sequence quality and rapidly increasing volumes of sequencing data. Combined with the fact that we for the last couple of years have seen a significant increase in the number of researchers involved in next generation sequencing projects, this creates a need for de novo assemblers which can both handle extremely large amounts of data and solve the complex task of constructing large contigs from short read data. Thus, a continuous development of de novo assemblers is crucial in order to benefit from the improved quality and quantity of sequencing data which are becoming available today.

Platform	Illumina MiSeq	454 FLX	Ion Torrent
E. coli strain	DH10B	0104:H4	DH10B
# Bases	2.18Gb	127Mb	990Mb
Avg. coverage	464x	23.9x	211x
Library type	Paired-end	Unpaired	Mate pair (62%) Unpaired (38%)
Insert size	300b	N/A	10kb
Avg. read length	151b	363b	140b

Table 1. E. coli sequencing data statistics

Benchmarks

Performance benchmarks have been executed for Escherichia coli, Arabidopsis thaliana and Homo sapiens for the assembly algorithm in CLC Assembly Cell 4.0. Normally, the ideal result of a de novo assembly is a few large contigs. A common way to measure assembly quality is, therefore, to compute statistics on the number and size of contigs which indicate how many reads the assembler is able to merge. Quality measurements for our benchmarks include, among others, the number of contigs produced, the total sum of bases in all contigs, and the N50 statistics computed by first summing the sizes of contigs in the order largest to smallest until the sum is $\geq \frac{1}{2} \sum|\text{contig}|$. The N50 is then the size of the contig that was last added to the sum.

Dataset	# contigs	Σ contig (Mb)	N50 (Kb)
Unpaired reads with default parameters			
Illumina	177	4.48	76.3
454 FLX	2,555	5.62	6.9
Ion Torrent	747	4.56	30.6
Use of paired information			
Illumina	127 (112)	4.49 (4.49)	88.3 (107.6)
454 FLX	N/A	N/A	N/A
Ion Torrent	716 (626)	4.57 (4.69)	32.2 (129.7)
Trimmed reads			
Illumina	113 (99)	4.48 (4.49)	97.0 (107.6)
454 FLX	1326	5.44	13.6
Ion Torrent	270 (136)	4.52 (4.66)	55.2 (2,950)
Word size			
Illumina, 23-23	113 (99)	4.48 (4.49)	97.0 (107.6)
454 FLX, 20-22	1320	5.43	15.1
Ion Torrent, 21-21	270 (136)	4.52 (4.66)	55.2 (2,950)
Bubble size			
Illumina, 50-50	113 (99)	4.48 (4.49)	97.0 (107.6)
454 FLX 50-300	1002	5.26	25.8
Ion Torrent 50-50	270 (136)	4.52 (4.66)	55.2 (2,950)

Table 2. Results of assemblies using different parameter settings for the CLC assembler. x=>y indicate that a parameter was changed from x to y. Where paired information is available, results for scaffolding are shown in parentheses.

De novo assembly of E. coli

To assess performance of the CLC assemblers on small datasets, we executed de novo assem-

CPU	i7 2720QM (laptop)	2x Xeon X5550	4x Xeon E7-4870
Cores	8	16	80
Threads used	8	16	40
Time used	5m 12s	4m 8s	2m 25s
Peak memory	129MB	287MB	778MB

Table 3. Time and memory used by the CLC assembler on three different systems when assembling Illumina MiSeq reads from E. coli DH10B. The numbers include scaffolding of the contigs.

Platform	Illumina GAII	Illumina GAIIx	Illumina GAIIx
# Bases	4.97Gb	3.21Gb	1.85Gb
Avg. coverage	42.2x	27.2x	15.7
Library type	Paired-end	Paired-end	Unpaired
Insert size	200b	400b	N/A
Avg. read length	36b	51b	51b

Table 4. A. thaliana Edi-0 sequencing data statistics

Assembler	CLC bio	SOAPdenovo
# Contigs	26,606 (13,164)	322,757
Σ contig (Mb)	106.2 (108.1)*	117.0*
N50 (Kb)	14.7 (35.0)	1.9

Table 5. De novo assembly of A. thaliana using CLC and SOAP de novo assemblers. Scaffolding results are shown in parentheses. *Minimum contig length for CLC and SOAP were 200b and 50b, respectively. Reference: Gan et al., Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature, 477(7365):419–423, 2011

CPU	i7 2720QM (laptop)	2x Xeon X5550	4x Xeon E7-4870
Cores	8	16	80
Threads used	8	16	40
Time used	30m 7s	17m 41s	13m 16s
Peak memory	2.8GB	2.8GB	3.2GB

Table 6. Performance of CLC Assembler on three different systems when assembling the A. thaliana Edi-0 genome using 10Gb of Illumina reads. All numbers include scaffolding of the contigs.

Platform	Illumina GAII
# Bases	135.3 Gb
Avg. coverage	43.7x
Library type	Paired-end
Insert type	300b
Avg. read length	101b

Table 7. H. sapiens sequencing data statistics

blies for two different strains of E. coli using high quality reads from three popular next generation sequencing platforms.

To optimize the assembly, we adjusted the parameters of the CLC assembler. Taking paired read information into account resulted in fewer and longer contigs for both the Illumina and Ion Torrent datasets. Paired reads also make it possible to perform scaffolding and in the case of the Ion Torrent dataset 716 contigs were reduced to 626 scaffolds with one scaffold covering 344Kb. Quality trimming also showed an impact on all of these datasets and by trimming the reads, the assembler was able to find more overlapping reads and merge several contigs. This resulted in larger N50 values. The small Illumina dataset was assembled on three different systems, including one laptop. On all systems, the assembly was completed in a few minutes using less than 1GB of memory.

De novo assembly of A. thaliana

The relatively small diploid genome of A. thaliana is one of the most well-studied model organisms in plant biology and, therefore, well suited for benchmarking the CLC de novo assembler on a plant genome.

To optimize the assembly, the automatically chosen word size of 24 was increased to 26 due to the high coverage while the bubble size was reduced to 40 due to the short read length. A total length of contigs of 90% of the reference genome was obtained. Time and memory usage for de novo assembly of the A. thaliana genome demonstrate that the CLC assembler is able to assemble fairly large genomes on standard consumer systems efficiently.

De novo assembly of H. sapiens

To assess the performance of the CLC assembler on a large genome, we assembled the genome of Hapmap individual NA18507. To improve assembly accuracy, the default parameters of word size and bubble size were increased from 26 to 32 and from 50 to 80, respectively. A total of 106 contigs were produced with a total length of 2.51Gb which corresponds to 87% of the known bases in the

# Contigs	1,079,610 (985,587)
Σ contig (Gb)	2.51 (2.52)
N50 (b)	5717 (6256)

Table 8. De novo assembly of H. sapiens using the CLC de novo assembler. Scaffolding results are shown in parentheses.

CPU	2x Xeon X5550	4x Xeon E7-4870
Cores	16	80
Threads used	16	40
Time used	11h 49m	7h 1m
Peak memory	45.3GB	46.9GB

Table 9. Performance of the CLC assembler on two different systems when assembling a human genome. The numbers include scaffolding of the contigs.

reference genome and 81% of the total reference genome length.

Assembly of this data-set demonstrated better efficiency in both time and memory consumption using the CLC assembler compared with other assembly algorithms such as e.g. SOAP and ABySS.

Conclusions

The computationally efficient algorithms used in the CLC assembler enable fast de novo assembly of high coverage bacterial and eukaryotic genomes using a laptop. Furthermore, an efficient parallelization scheme ensures full utilization of multicore processors, thus enabling de novo assemblies of the human genome to be completed in seven hours using a single computer. The benchmark results also indicate that the accuracy of the CLC assembler is similar to or better than other state of the art de novo assemblers which are capable of handling the massive amounts of short read data produced by modern sequencing technologies.

Both CLC Genomics Workbench and CLC Genomics Server integrate powerful bioinformatics tools under a unified graphical user interface. In contrast, CLC Assembly Cell is a high-performance application dedicated to users versed in operating software at the command line level. It is the purpose of both product lines to solve complex bioinformatics tasks, amongst others de novo assemblies and read mappings.

The CLC de novo assembly algorithm included in CLC Genomics Workbench, CLC Genomics Server, and CLC Assembly Cell demonstrates the ability to produce accurate assemblies extremely fast using both standard consumer hardware and server systems.