Shasta assembly summary

Shasta version

Shasta Release 0.11.1

Reads used in this assembly

Read representation	1 (RLE)
Minimum read length	500
Number of reads	7990424
Number of read bases	90036879397
Average read length	11268
Read N50	16495
Number of run-length encoded bases	63668139720
Average length ratio of run-length encoded sequence over raw sequence	0.7071
Number of reads flagged as palindromic by self alignment	294011
Number of reads flagged as chimeric	165850

- Reads discarded on input are not included in the above table (see below).
- See ReadLengthHistogram.csv and Binned-ReadLengthHistogram.csv for details of the read length distribution of reads used in this assembly.

Reads discarded on input

	Reads	Bases
Reads discarded on input because they contained invalid bases	0	0
Reads discarded on input because they were too short	188734	60899277
Reads discarded on input because they contained repeat counts greater than 255	7215	47948757
Reads discarded on input, total	195949	108848034
Fraction of reads discarded on input over total present in input files	0.02394	0.001207

- Base counts in the above table are raw sequence bases.
- Here and elsewhere, "raw" refers to the original read sequence, as opposed to run-length encoded sequence.

Marker *k*-mers

Length k of k-mers used as markers	14
Total number of k-mers	6377292
Number of k-mers used as markers	638031
Fraction of k-mers used as markers	0.1

- In the above table, all k-mer counts only include run-length encoded k-mers, that is, k-mers without repeated bases.

Markers

Total number of markers on all reads, one strand	6363005569
Total number of markers on all reads, both strands	12726011138
Average number of markers per raw base	0.07067
Average number of markers per run-length encoded base	0.09994
Average base offset between markers in raw sequence	14.15
Average base offset between markers in run-length encoded sequence	10.01
Average base gap between markers in run-length encoded sequence	-3.994

Alignments

Number of alignment candidates found by the LowHash algorithm	151822487
Number of good alignments	138802883
Number of good alignments kept in the read graph	32171192

Alignment criteria actually used for creation of the read graph

minAlignedMarkerCount	55
minAlignedFraction	0.545
maxSkip	19
maxDrift	9
maxTrim	54

Read graph

Number of vertices	15980848
Number of edges	64342384

- The read graph contains both strands. Each read generates two vertices.
- Isolated reads in the read graph don't contribute to the assembly. See the table below for a summary of isolated reads in the read graph. Each isolated read corresponds to two isolated vertices in the read graph, one for each strand.

	Reads	Bases
Isolated reads	1828333	12370926328
Non-isolated reads	6162091	77665953069
Isolated reads fraction	0.2288	0.1374
Non-isolated reads fraction	0.7712	0.8626

Marker graph

Total number of vertices	394171532
Total number of edges	842323402
Number of vertices that are not isolated after edge removal	378597526
Number of edges that were not removed	378593900

- The marker graph contains both strands.

Assembly graph

Number of vertices	16400
Number of edges	10126
Number of edges assembled	5063

- The assembly graph contains both strands.

Assembled segments ("contigs")

Number of segments assembled	5063
Total assembled segment length	2699573097
Longest assembled segment length	60846000
Assembled segments N_{50}	13229089

- Shasta uses GFA terminology (*segment* instead of the most common *contig*). A contiguous section of assembled sequence can consist of multiple segments, for example in the presence of heterozygous bubbles.
- See AssemblySummary.csv for lengths of assembled segments.

Performance

Elapsed time (seconds)	5.576e+04
Elapsed time (minutes)	929.4
Elapsed time (hours)	15.49
Average CPU utilization	0.2004
Peak virtual memory utilization (bytes)	577094090752
Number of threads used	32
Total number of virtual CPUs available	32
Total physical memory available (bytes)	405556690944