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	34443838
Number of read bases	386152292706
Average read length	11211
Read N50	14925
Number of run-length encoded bases	273010252845
Average length ratio of run-length encoded sequence over raw sequence	0.707
Number of reads flagged as palindromic by self alignment	470236
Number of reads flagged as chimeric	287230

- 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	445327	142706755
Reads discarded on input because they contained repeat counts greater than 255	41790	256426117
Reads discarded on input, total	487117	399132872
Fraction of reads discarded on input over total present in input files	0.01395	0.001033

- 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	27291003394
Total number of markers on all reads, both strands	54582006788
Average number of markers per raw base	0.07067
Average number of markers per run-length encoded base	0.09996
Average base offset between markers in raw sequence	14.15
Average base offset between markers in run-length encoded sequence	10
Average base gap between markers in run-length encoded sequence	-3.996

Alignments

Number of alignment candidates found by the LowHash algorithm	54017135
Number of good alignments	50469265
Number of good alignments kept in the read graph	22224669

Alignment criteria actually used for creation of the read graph

minAlignedMarkerCount	115
minAlignedFraction	0.595
maxSkip	19
maxDrift	10
maxTrim	49

Read graph

Number of vertices	68887676
Number of edges	44449338

- 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	27854343	281554887510
Non-isolated reads	6589495	104597405196
Isolated reads fraction	0.8087	0.7291
Non-isolated reads fraction	0.1913	0.2709

Marker graph

Total number of vertices	416896004
Total number of edges	832781336
Number of vertices that are not isolated after edge removal	397307762
Number of edges that were not removed	397157392

- The marker graph contains both strands.

Assembly graph

Number of vertices	329894
Number of edges	179132
Number of edges assembled	89566

- The assembly graph contains both strands.

Assembled segments ("contigs")

Number of segments assembled	89566
Total assembled segment length	3046328432
Longest assembled segment length	3679915
Assembled segments N_{50}	58281

- 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)	7.787e+05
Elapsed time (minutes)	1.298e+04
Elapsed time (hours)	216.3
Average CPU utilization	0.02593
Peak virtual memory utilization (bytes)	1656296235008
Number of threads used	32
Total number of virtual CPUs available	32
Total physical memory available (bytes)	405556690944