Shasta assembly summary

Shasta version

Shasta Release 0.11.1

Reads used in this assembly

Read representation	1 (RLE)
Minimum read length	50000
Number of reads	713397
Number of read bases	51921430407
Average read length	72780
Read N50	71805
Number of run-length encoded bases	35672149056
Average length ratio of run-length encoded sequence over raw sequence	0.687
Number of reads flagged as palindromic by self alignment	6048
Number of reads flagged as chimeric	17484

- 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	1626	68963182
Reads discarded on input because they contained repeat counts greater than 255	1900	141556339
Reads discarded on input, total	3526	210519521
Fraction of reads discarded on input over total present in input files	0.004918	0.004038

- 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	10
Total number of *k*-mers	78732
Number of *k*-mers used as markers	7764
Fraction of *k*-mers used as markers	0.0986

- 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	3546755436
Total number of markers on all reads, both strands	7093510872
Average number of markers per raw base	0.06831
Average number of markers per run-length encoded base	0.09943
Average base offset between markers in raw sequence	14.64
Average base offset between markers in run-length encoded sequence	10.06
Average base gap between markers in run-length encoded sequence	0.05769

Alignments

Number of alignment candidates found by the LowHash algorithm	22680402
Number of good alignments	16245628
Number of good alignments kept in the read graph	5162087

Alignment criteria actually used for creation of the read graph

minAlignedMarkerCount	195
minAlignedFraction	0.255
maxSkip	66
maxDrift	22
maxTrim	60

Read graph

Number of vertices	1426794
Number of edges	10324174

- 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	63201	4413425486
Non-isolated reads	650196	47508004921
Isolated reads fraction	0.08859	0.085
Non-isolated reads fraction	0.9114	0.915

Marker graph

Total number of vertices	114007008

Total number of edges	131414388
Number of vertices that are not isolated after edge removal	0
Number of edges that were not removed	0

- The marker graph contains both strands.

Phased assembly statistics

	Length	N_{50}
Bubble chains	337113271	512604
Diploid sequence (per haplotype)	169712323	253023
Haploid sequence	169111147	425483
Total sequence assembled in bubble chains, per haplotype	338823470	
Sequence outside bubble chains	565290326	

Number of bubbles that describe a single SNP (transition)	10360
Number of bubbles that describe a single SNP (transversion)	7252
Number of bubbles that describe a single SNP (total)	17612
Transition/transversion ratio for bubbles that describe a single SNP	1.429
Number of bubbles that describe indels or more than one SNP	96730

Performance

Elapsed time (seconds)	3950
Elapsed time (minutes)	65.83
Elapsed time (hours)	1.097
Average CPU utilization	0.2498
Peak virtual memory utilization (bytes)	273328578560
Number of threads used	192
Total number of virtual CPUs available	192
Total physical memory available (bytes)	2163841273856