Shasta assembly summary

Shasta version

Shasta Release 0.8.0

Reads used in this assembly

Minimum read length	10000
Number of reads	7912275
Number of raw sequence bases	183287165424
Average read length (for raw read sequence)	23164
Read N50 (for raw read sequence)	26270
Number of run-length encoded bases	128549251212
Average length ratio of run-length encoded sequence over raw sequence	0.7014
Number of reads flagged as palindromic by self alignment	176217
Number of reads flagged as chimeric	203204

- Here and elsewhere, "raw" refers to the original read sequence, as opposed to run-length encoded sequence.
- 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	10212191	40740451958
Reads discarded on input because they contained repeat counts greater than 255	558	12665350
Reads discarded on input because they had quality scores indicative of palindromic sequence	0	0
Reads discarded on input, total	10212749	40753117308
Fraction of reads discarded on input over total present in input files	0.5635	0.1819

- 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	12869340013
Total number of markers on all reads, both strands	25738680026
Average number of markers per raw base	0.07021
Average number of markers per run-length encoded base	0.1001
Average base offset between markers in raw sequence	14.24
Average base offset between markers in run-length encoded sequence	9.989
Average base gap between markers in run-length encoded sequence	-4.011

- Here and elsewhere, "raw" refers to the original read sequence, as opposed to run-length encoded sequence.

Alignments

Number of alignment candidates found by the LowHash algorithm	78393966
Number of good alignments	60518063
Number of good alignments kept in the read graph	19761171

Read graph

Number of vertices	15824550
Number of edges	39522342

- 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	3472875	72503520963
Non-isolated reads	4439400	110783644461
Isolated reads fraction	0.4389	0.3956
Non-isolated reads fraction	0.5611	0.6044

Marker graph

Total number of vertices	415864158
Total number of edges	654994186
Number of vertices that are not isolated after edge removal	367116594
Number of edges that were not removed	367077854

- The marker graph contains both strands.

Assembly graph

Number of vertices	102054
Number of edges	62308
Number of edges assembled	31154

- The assembly graph contains both strands.

Assembled segments ("contigs")

Number of segments assembled	31154
Total assembled segment length	2703495682
Longest assembled segment length	8971914
Assembled segments N_{50}	147996

- 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)	1.235e+04
Elapsed time (minutes)	205.8
Elapsed time (hours)	3.43
Average CPU utilization	0.3029
Peak Memory utilization (bytes)	828513337344