

# Shasta assembly summary

## Shasta version

Shasta Release 0.11.1

## Reads used in this assembly

Read representation	1 (RLE)
Minimum read length	16724
Number of reads	6844513
Number of read bases	160011478004
Average read length	23378
Read N50	22807
Number of run-length encoded bases	113096381369
Average length ratio of run-length encoded sequence over raw sequence	0.7068
Number of reads flagged as palindromic by self alignment	134778
Number of reads flagged as chimeric	368600

- 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	28044652	226283521457
Reads discarded on input because they contained repeat counts greater than 255	41790	256426117
Reads discarded on input, total	28086442	226539947574
Fraction of reads discarded on input over total present in input files	0.8041	0.5861

- 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 <i>k</i> of <i>k</i> -mers used as markers	14
Total number of <i>k</i> -mers	6377292
Number of <i>k</i> -mers used as markers	638031
Fraction of <i>k</i> -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	11314942766
Total number of markers on all reads, both strands	22629885532
Average number of markers per raw base	0.07071
Average number of markers per run-length encoded base	0.1
Average base offset between markers in raw sequence	14.14
Average base offset between markers in run-length encoded sequence	9.995
Average base gap between markers in run-length encoded sequence	-4.005

## Alignments

Number of alignment candidates found by the LowHash algorithm	239752029
Number of good alignments	191002520
Number of good alignments kept in the read graph	30404097

## Alignment criteria actually used for creation of the read graph

minAlignedMarkerCount	115
minAlignedFraction	0.525
maxSkip	25
maxDrift	11
maxTrim	62

## Read graph

Number of vertices	13689026
Number of edges	60808194

- 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	733444	17306555323
Non-isolated reads	6111069	142704922681
Isolated reads fraction	0.1072	0.1082
Non-isolated reads fraction	0.8928	0.8918

## Marker graph

Total number of vertices	392345650
Total number of edges	1071645528
Number of vertices that are not isolated after edge removal	376440588
Number of edges that were not removed	376440596

- The marker graph contains both strands.

## Assembly graph

Number of vertices	10536
Number of edges	6696
Number of edges assembled	3348

- The assembly graph contains both strands.

## Assembled segments ("contigs")

Number of segments assembled	3348
Total assembled segment length	2690380736
Longest assembled segment length	40738402
Assembled segments N <sub>50</sub>	7724828

- 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)	4.65e+05
Elapsed time (minutes)	7750
Elapsed time (hours)	129.2
Average CPU utilization	0.09569
Peak virtual memory utilization (bytes)	932598161408
Number of threads used	32
Total number of virtual CPUs available	32
Total physical memory available (bytes)	135075577856