

Shasta 0.13.0

Konstantinos Kyriakidis
UC Santa Cruz

Shasta “Mode 3” assembly*

- Released in preliminary form with Shasta 0.12.0 and further improved in this 0.13.0 software release.
- Despite known issues (to be improved on in future releases), produces useful and accurate phased assemblies using high accuracy nanopore reads from the ONT London Calling 2024 data release (https://labs.epi2me.io/lc2024_t2t/)
- Like previous Shasta releases, uses markers, MinHash, read graph, marker graph.
- Final sequence assembly is new.
 - Uses the marker graph to locate features that are unique to a single location+haplotype in the assembly.
 - “Read following” on these unique features.
 - Then uses local assemblies to assemble sequences between unique features.
- Invoked with *--config Nanopore-r10.4.1_e8.2-400bps_sup-Raw-Sep2024* for raw reads and *--config Nanopore-r10.4.1_e8.2-400bps_sup-Herro-Sep2024* for HERRO corrected reads
- Sequence assembly for a human genome takes 2-5 hours on a machine of appropriate size, depending on coverage.
- Memory requirement is currently 6 bytes per input base.
 - A 1 TB machine can run a human assembly at 50x.

* Shasta detailed Computational Methods:

<https://paoloshasta.github.io/shasta/ComputationalMethods.html>

New ONT Q25 Chemistry used for benchmarking

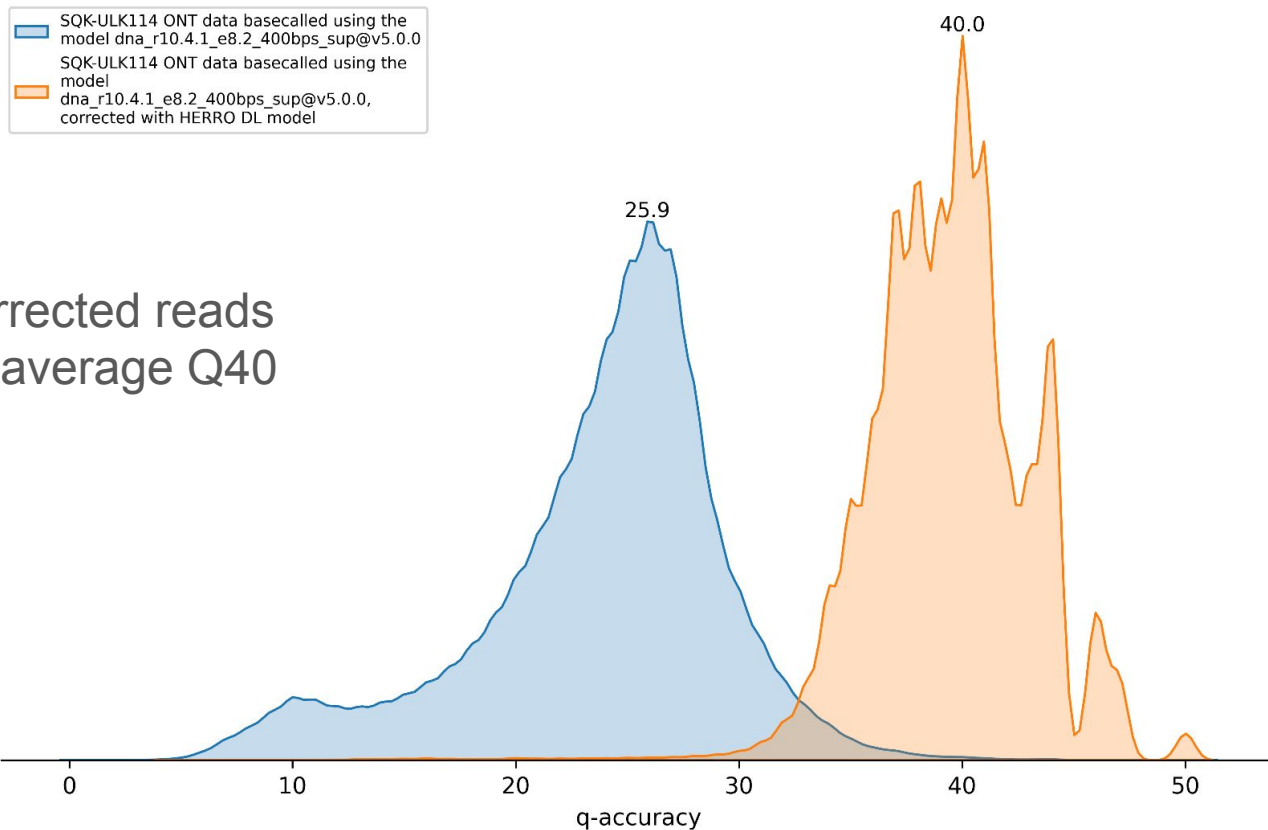
For the benchmarking of the assemblies we used the following data:

- Two Ultra Long (SQK-ULK114) runs, basecalled using the model dna_r10.4.1_e8.2_400bps_sup@v5.0.0
- The HERRO corrected data of the above two runs

Both datasets are part of “London Calling 2024: a Nanopore-only telomere-to-telomere (T2T) assembly dataset” and are available here:

https://labs.epi2me.io/lc2024_t2t/

New ONT Q25 Chemistry used for benchmarking



HERRO corrected reads
achieve an average Q40

Comparisons between Shasta v0.13.0 and Verkko v2.1

Shasta assemblies were run using the following commands:

```
shasta --config Nanopore-r10.4.1_e8.2-400bps_sup-Raw-Sep2024 --threads ${THREADS} --input  
${RAW_READS} --assemblyDirectory ${OUTPUT_PATH}
```

```
shasta --config Nanopore-r10.4.1_e8.2-400bps_sup-Herro-Sep2024 --threads ${THREADS} --input  
${HERRO_CORRECTED_READS} --assemblyDirectory ${OUTPUT_PATH}
```

Verkko assembly was run using both the HERRO corrected reads and the RAW reads as input (the recommended approach) using the following command:

```
verkko -d ${OUTPUT_PATH} --hifi ${HERRO_CORRECTED_READS} --nano ${RAW_READS}
```

Assembly metrics

The “London Calling 2024: a Nanopore-only telomere-to-telomere (T2T) assembly” dataset used in these benchmarks has **coverage ~45x**.

	HG002 v1.1	SHASTA v0.13.0 RAW	SHASTA v0.13.0 HERRO	VERKKO v2.1 RAW + HERRO
Total sequence assembled (Mb)	5999	5921	5957	6116
N50 (Mb)	147	42.1	62.9	40.5

Comparison on chr18_MATERNAL

All following comparisons used mappings to the HG002 v1.1 reference haplotypes.

chr18_Maternal was assembled completely by all three assemblies. Verkko v2.1 managed to assemble it T2T while Shasta v0.13.0 missed a few hundred telomere bases.

chr18 Maternal	SHASTA v0.13.0	SHASTA v0.13.0	VERKKO v2.1
Datasets Used	RAW	HERRO	RAW + HERRO
Mismatches	40	14	92
Insertions	1504	1536	3528
Deletions	7698	5103	5280

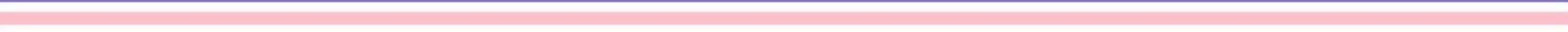



Note: The numbers reported are number of bases (that is, longer errors count more)

Comparison of assembled contigs

Alignments to chr4_PATERNAL

This reference segment is 192384391 bases long and has 4 alignments.

Alignments to chr4_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr4_PATERNAL








chr4_PATERNAL	
contig-0001359	
contig-0000239	
contig-0000203	
contig-0000200	

Alignments to chr4_MATERNAL

This reference segment is 191670063 bases long and has 12 alignments.

VERKKO v2.1 (RAW + HERRO)

Alignments to chr4_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr4_MATERNAL

chr4_MATERNAL	
contig-0001061	
contig-0001058	
contig-0001059	
contig-0001060	
contig-0000627	
contig-0001347	
contig-0001348	
contig-0000424	
contig-0001346	
contig-0000204	
contig-0000199	
contig-0000201	

Comparison of assembled contigs

Alignments to chr4_PATERNAL

This reference segment is 192384391 bases long and has 3 alignments.

Alignments to chr4_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr4_PATERNAL

chr4_PATERNAL	
4-3-0-0-P0	
4-0-0-0-P0	
4-2-0-0-P0	

Alignments to chr4_MATERNAL

This reference segment is 191670063 bases long and has 3 alignments.

Alignments to chr4_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr4_MATERNAL

chr4_MATERNAL	
4-4-0-0-P0	
29-0-0-0-P0	
4-1-0-0-P0	

SHASTA v0.13.0 (HERRO)

Alignments to chr5_PATERNAL

This reference segment is 188875663 bases long and has 26 alignments.

VERKKO (RAW + HERRO)

Alignments to chr5_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr5_PATERNAL

chr5_PATERNAL	
contig-0001317	
contig-0000370	
contig-0000811	
contig-0000830	
contig-0001427	
contig-0001429	
contig-0000371	
contig-0001029	
contig-0000817	
contig-0000803	
contig-0000822	
contig-0001428	
contig-0001430	
contig-0001028	
contig-0001027	
contig-0000131	
contig-0000072	
contig-0000252	
contig-0000497	
contig-0000259	
contig-0000243	
contig-0000464	
contig-0000474	
contig-0000297	
contig-0000274	
contig-0000273	

Alignments to chr5_MATERNAL

This reference segment is 183262196 bases long and has 11 alignments.

Alignments to chr5_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr5_MATERNAL




chr5_MATERNAL	
contig-0001341	
contig-0000070	
contig-0000492	
contig-0000073	
contig-0000482	
contig-0000295	
contig-0000296	
contig-0000717	
contig-0000275	
contig-0000271	
contig-0000272	

Alignments to chr5_PATERNAL

This reference segment is 188875663 bases long and has 25 alignments.

SHASTA v0.13.0 (HERRO)

Alignments to chr5_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr5_PATERNAL

chr5_PATERNAL	
5-14-0-0-P0	
5-16-0-0-P1	
5-16-1-1-P2	
5-10-0-0-P0	
5-18-0-0-P2	
5-18-1-0-P1	
5-9-0-0-P0	
5-4-0-0-P0	
5-7-0-0-P0	
5-17-0-0-P1	
5-17-1-0-P2	
5-17-2-0-P1	
5-17-3-1-P2	
5-5-0-0-P0	
5-19-0-0-P1	
5-19-1-0-P2	
5-17-1-1-P2	
5-17-3-0-P2	
5-17-4-0-P1	
5-6-0-0-P0	
5-18-0-1-P2	
5-19-1-1-P2	
5-8-0-0-P0	
5-15-0-0-P0	
5-11-0-0-P0	

Alignments to chr5_MATERNAL

This reference segment is 183262196 bases long and has 3 alignments.

Alignments to chr5_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr5_MATERNAL

chr5_MATERNAL	
5-13-0-0-P0	
5-2-0-0-P0	
5-12-0-0-P0	

Alignments to chrX_MATERNAL

This reference segment is 154341406 bases long and has 1 alignments.

Alignments to chrX_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chrX_MATERNAL

chrX_MATERNAL	
contig-0000590	

Alignments to chrX_MATERNAL sorted by begin position in chrX_MATERNAL

Reference segment			Assembled segment						Matching base count	Alignment length	Mapping quality	Number			Rate				Q (dB)			
Begin	End	Aligned Length	Name	Length	Strand	Begin	End	Aligned length				Mismatch	Insert	Delete	Mismatch	Insert	Delete	Identity	Mismatch	Insert	Delete	Identity
0	154341406	154341406	contig-0000590	154383146	+	2869	154381354	154378485	154330447	154389401	60	43	47995	10916	2.79e-07	3.11e-04	7.07e-05	0.9996	65.6	35.1	41.5	34.2

Alignments to chrY_PATERNAL

This reference segment is 62425491 bases long and has 1 alignments.

VERKKO v2.1 (RAW + HERRO)

Alignments to chrY_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chrY_PATERNAL

chrY_PATERNAL	
contig-0000910	

Alignments to chrY_PATERNAL sorted by begin position in chrY_PATERNAL

Reference segment			Assembled segment						Matching base count	Alignment length	Mapping quality	Number			Rate				Q (dB)			
Begin	End	Aligned Length	Name	Length	Strand	Begin	End	Aligned length				Mismatch	Insert	Delete	Mismatch	Insert	Delete	Identity	Mismatch	Insert	Delete	Identity
1	62425491	62425490	contig-0000910	62437469	+	3212	62434082	62430870	62423151	62433195	60	14	7705	2325	2.24e-07	1.23e-04	3.72e-05	0.9998	66.5	39.1	44.3	37.9

Shasta performs suboptimally on sex chromosomes

SHASTA v0.13.0 (HERRO)

chrY_PATERNAL	
11-125-2-0-P1	
11-125-1-0-P2	
11-125-1-1-P2	
11-125-0-0-P1	
11-69-0-0-P0	
11-76-0-0-P0	
11-93-0-0-P0	
11-71-0-0-P0	
11-92-0-0-P0	
11-87-0-0-P0	
11-120-0-0-P0	
11-77-0-0-P0	
11-104-0-0-P0	
11-89-0-0-P0	
11-90-0-0-P0	
11-88-0-0-P0	
11-118-0-0-P1	
11-68-0-0-P0	
11-73-0-0-P0	
11-103-0-0-P0	
11-94-0-0-P0	
11-78-0-0-P0	
11-102-0-0-P0	
11-84-0-0-P0	
11-99-0-0-P0	
11-95-0-0-P0	
11-97-0-0-P0	
11-100-0-0-P0	
11-72-0-0-P0	
11-101-0-0-P0	
11-98-0-0-P0	
11-96-0-0-P0	
11-124-0-0-P1	
11-124-1-1-P2	
11-124-1-0-P2	
11-124-2-0-P1	

Alignments to chrX_MATERNAL

This reference segment is 154341406 bases long and has 7 alignments.

Alignments to chrX_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chrX_MATERNAL

chrX_MATERNAL	
11-122-4-0-P1	
11-122-3-0-P2	
11-122-3-1-P2	
11-122-2-0-P1	
11-122-1-0-P2	
11-122-0-0-P1	
11-74-0-0-P0	