

Shasta v0.14.0

Konstantinos Kyriakidis

Paten lab
Miga lab
UC Santa Cruz

**Haplotype-resolved de novo assembly is the
ultimate solution to the study of sequence
variations in a genome**

The primary goal of Shasta is to produce the
**most accurate and complete phased
assemblies**

and do it

fast

How fast?

Sequence assembly for a human genome takes **2-5 hours** on a machine of appropriate size, **depending on coverage**.

~40x coverage takes about **3 hours** to run

Memory requirement is currently **6 bytes per input base**.

A **1 TB** machine can run a **human assembly at 50x**.

New ONT Q26 Chemistry

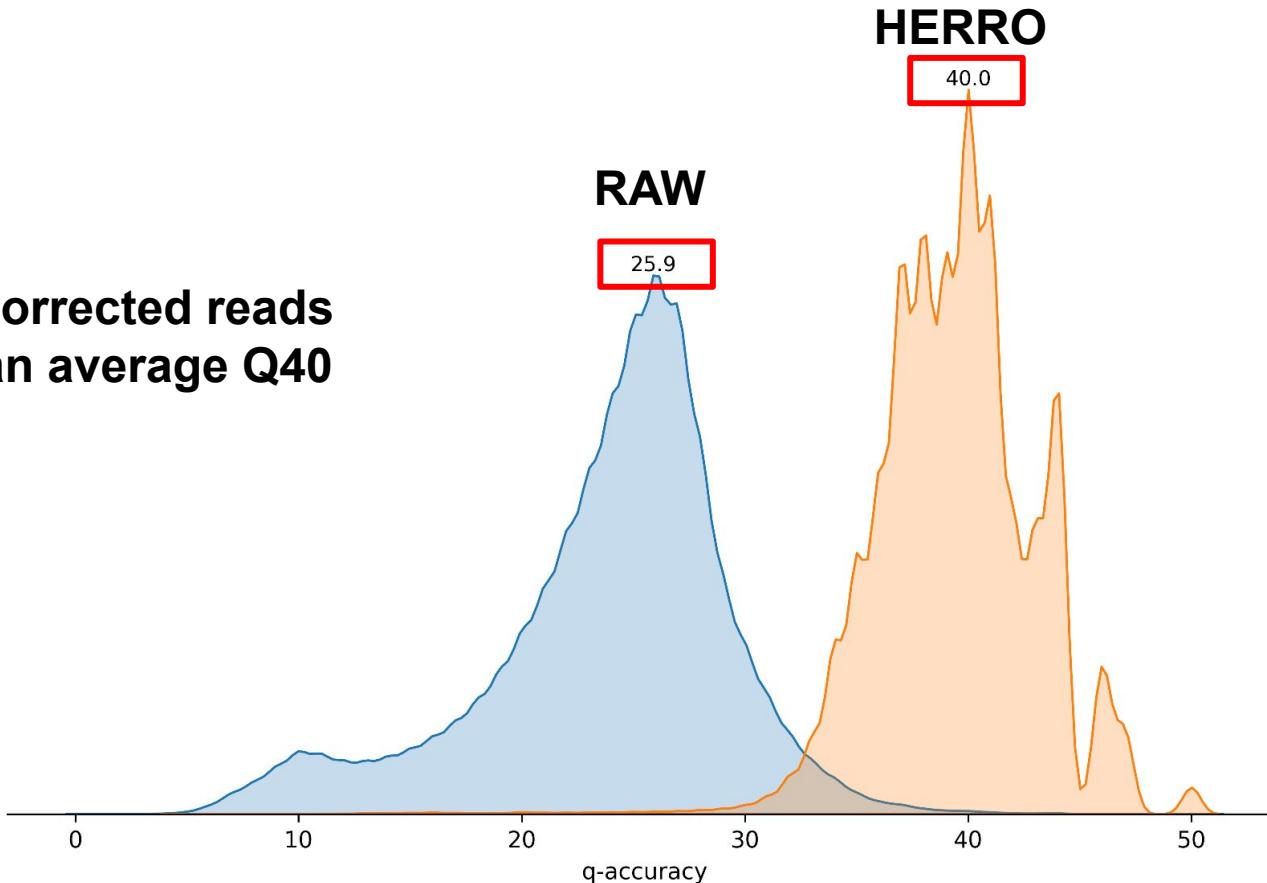
For optimizing and benchmarking Shasta we used the following data:

- **Two Ultra Long (SQK-ULK114) runs**, basecalled using the model dna_r10.4.1_e8.2_400bps_sup@v5.0.0
 - ~100Kb N50
 - ~40x coverage
- 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 Q26 Chemistry

**HERRO corrected reads
achieve an average Q40**



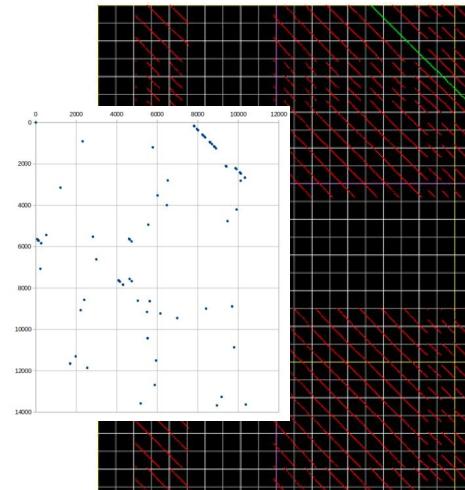
Shasta, simplified

(1) Represent reads as “markers”

Shasta, simplified

(1) Represent reads as “markers”

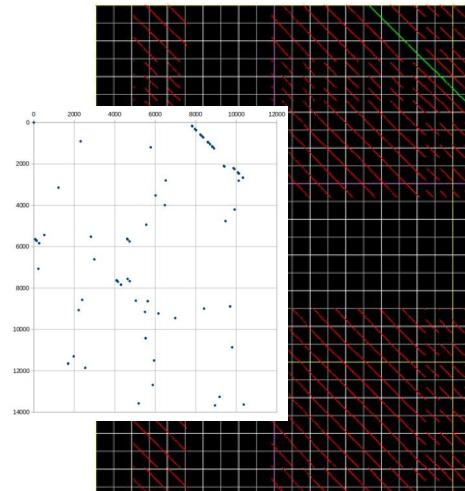
(2) MinHash/align reads as marker sequences



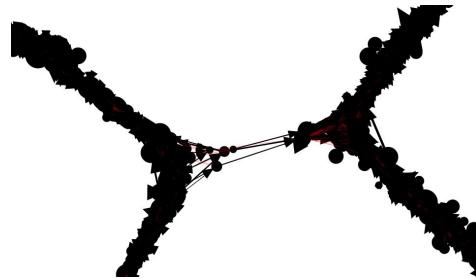
Shasta, simplified

(1) Represent reads as “markers”

(2) MinHash/align reads as marker sequences



(3) Construct read overlap graph to prune overlaps



Shasta, simplified

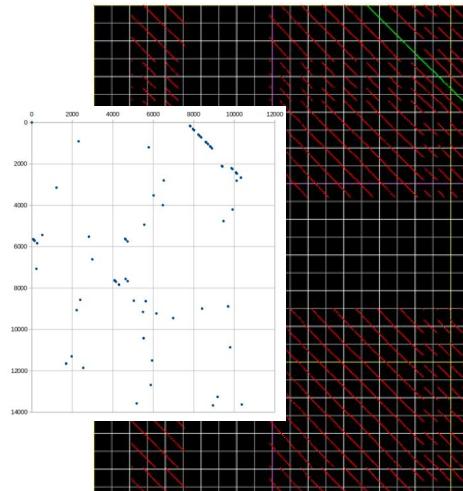
(1) Represent reads as “markers”

```

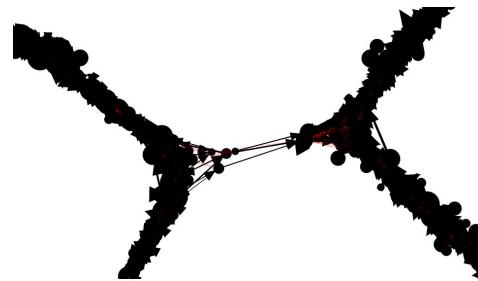
560      570      580      590      600      610      620      630      640      650      660
. . . . + . . . . + . . . | . . . + . . . | . . . + . . . | . . . + . . . | . . . + . . . | . . .
11221111321112213111113121112111211131121122111112112111241522531412221112111131131111
ATATCATCGATGAGTACAGCTGACTATCACTCATATCAGACTACTGACATGTGACTCATAGTGCTACTAGTCAGTCTATGTGTATGTGAT/
TATCATCGCA ATCTAGTAC GACTACTGAC TCATAGTGCT CTAGTCAGTC ATGTGTATGT
GCATGATCTG TGACATGTGA CATAGTGCTA AGTCTATGTG
CTGAGTACAG TGAGTACAGC TGTGTATGTG GTGTATGTGT

```

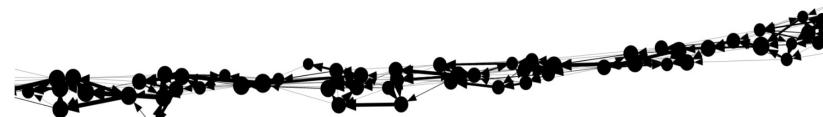
(2) MinHash/align reads as marker sequences



(3) Construct read overlap graph to prune overlaps



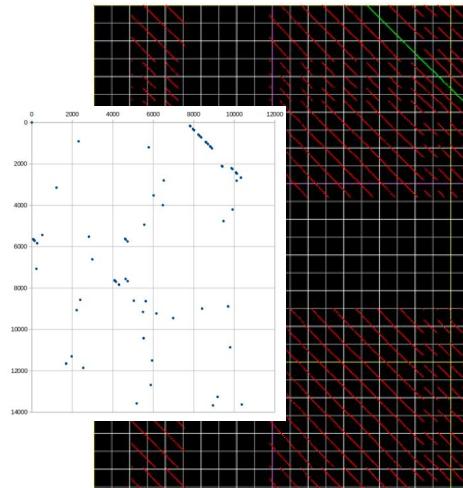
(4) Construct marker graph (MG)
representing aligned reads



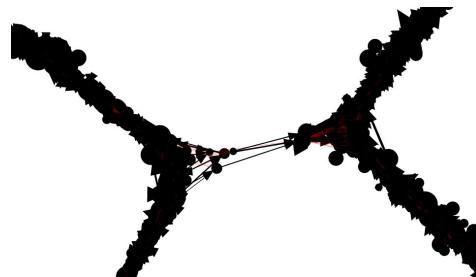
Shasta, simplified

(1) Represent reads as “markers”

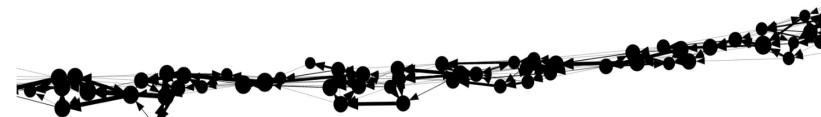
(2) MinHash/align reads as marker sequences



(3) Construct read overlap graph to prune overlaps

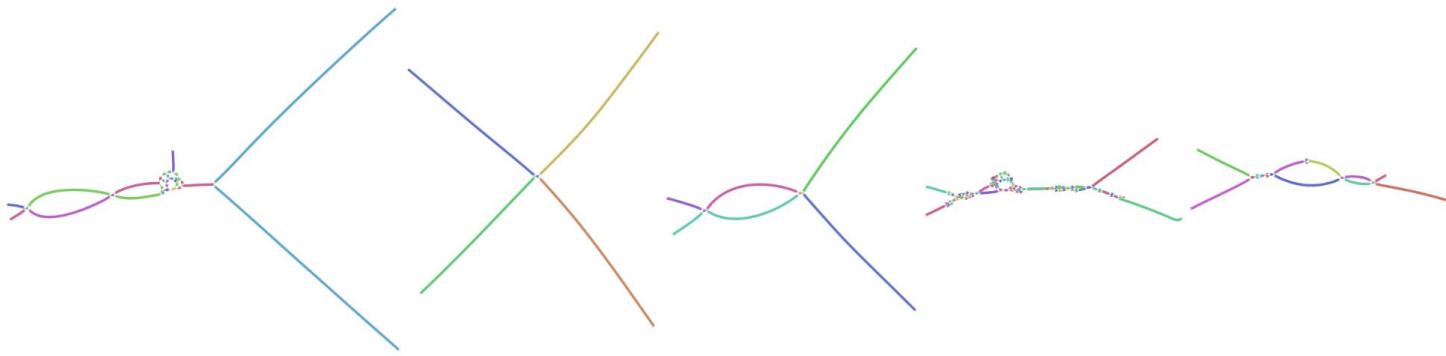


(4) Construct marker graph (MG) representing aligned reads

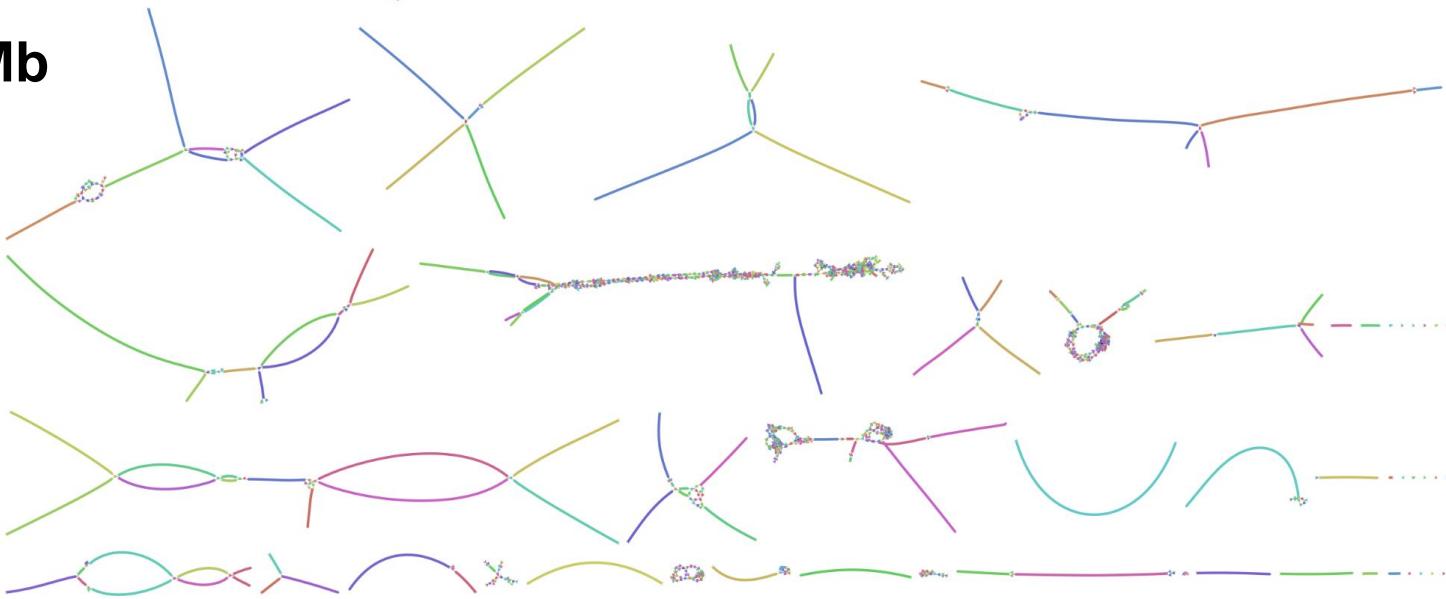


(5) NEW: Trace haplotypes in MG to assemble sequence - aka "Mode 3"

SHASTA v0.13.0 (HERRO) September 2024



N50 ~43Mb



SHASTA v0.14.0 (HERRO) January 2025

N50 ~80Mb



In case you are wondering about the N50...

Hifiasm v0.24.0 using HERRO/Dorado corrected Q26 ONT UL
reads

~98Mb N50

Shasta v0.14.0 using HERRO/Dorado corrected Q26 ONT UL
reads

~80Mb N50

Verkko v2.2.1 using HERRO/Dorado corrected Q26 ONT UL reads

~40Mb N50

Does this mean that Hifiasm perform better than
Shasta and Verkko?

Or, to phrase it differently, does this mean that
Verkko has the worst performance?

N50 is a “weak” metric if it is not paired with metrics
on how accurate the assembled contigs are

How to best evaluate/improve the assembly methods?

The data used to optimize and benchmark Shasta come from HG002 samples released by ONT in May 2024, in both raw and error corrected form

Luckily, we have the most accurate HG002 reference genome in our toolkit

The screenshot shows the GitHub repository page for HG002. The repository is public and has 1 branch, 0 tags, 42 commits, and 12 forks. The README file describes the Telomere-to-telomere consortium HG002 "Q100" project. It mentions the Human Genome Reference Consortium and the Genome in a Bottle Consortium, and details the assembly process using Verko and manual patching.

Code | Issues 1 | Pull requests | Actions | Projects | Security | Insights

HG002 Public

Watch 12 · Fork 5 · Starred 106

main · Branch · Tags

Go to file · Add file · Code

About

A complete diploid human genome

Readme · Activity · Custom properties · 106 stars · 12 watching · 5 forks · Report repository

Releases

No releases published

Packages

No packages published

Contributors 4

nhansen Nancy F. Hansen · aphillipy Adam Phillippy · skoren Sergey Koren · brianwalenz Brian Walenz

README

Telomere-to-telomere consortium HG002 "Q100" project

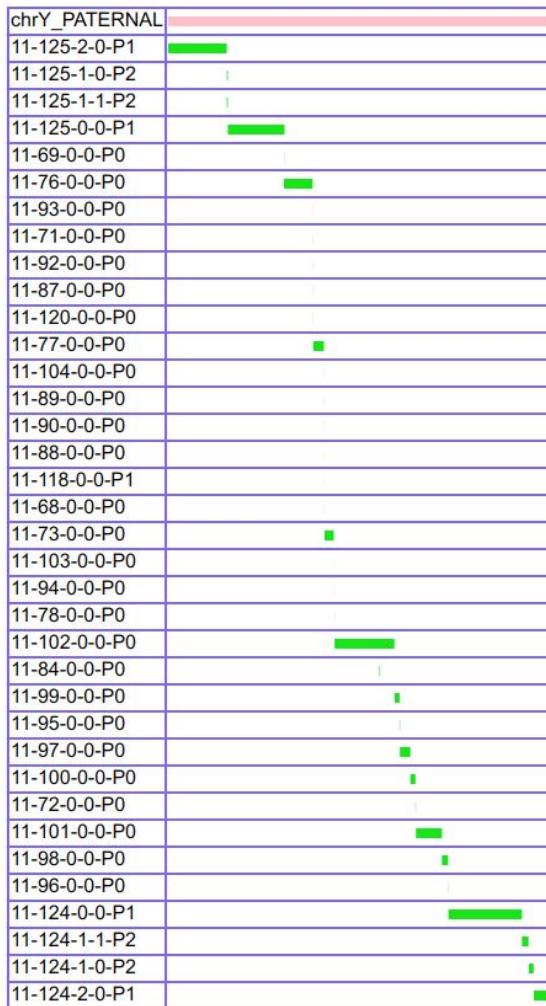
The [Telomere-to-Telomere Consortium](#), along with the [Human Pangenome Reference Consortium](#) and the [Genome in a Bottle Consortium](#), have sequenced, assembled and polished the [HG002](#) (also known as GM24385 and huAA53E0) cell line. The ultimate goal of this effort is to create a "genome benchmark" for the HG002 reference material that covers all bases of the diploid genome and is perfectly accurate. Hence, the "Q100" project nickname, which refers to a Phred quality score of 1 error per 10 billion bases.

Current benchmarks are typically defined as a list of variants called against a reference genome such as GRCh38. This is problematic in that the GRCh38 reference is incomplete, and there may be regions of the benchmark genome (e.g. HG002) that do not reliably align to the reference. A more natural and comprehensive benchmark representation is the complete sequence of the genome itself, i.e. a "genome benchmark" as opposed to a "variant benchmark". Assembling the complete HG002 genome is our first step towards creating such a genome benchmark, and next steps will include the development of tools and metrics for evaluating de novo assemblies and variant calls against the benchmark.

The initial assembly used for this project was performed using HiFi data available from the HPRC, as well as ONT data available from the HPRC and GIBAB. The [Verko](#) assembler was used, followed by manual assignment of nodes to chromosomes, ONT-based patching to resolve HiFi coverage gaps, manual resolution of tangles, and Strand-Seq and Hi-C based scaffolding across the rDNA arrays. The v0.7 assembly release then underwent three rounds of extensive polishing, patching, and validation to produce the v1.1 release. Although the v1.1 assembly does contain gaps (scaffolded stretches of N's) for nine out of ten of the rDNA arrays, it is otherwise T2T ("telomere to telomere"), has nearly perfect haplotype phasing, and has an estimated consensus error rate of less than 1 per 10 million bases (merqury assigned QV of Q78). Work on the annotation, characterization, and correction of any remaining errors in the v1.1 assembly will continue as sequencing technology improves and any additional errors are identified. This will include the eventual inclusion and finishing of the rDNA arrays. If you identify any errors in the latest assembly, please raise an issue with all relevant evidence and information on the [associated issues repository](#).

Let's have a closer look

chrY



Contig mappings against the T2T HG002v1.1 reference genome

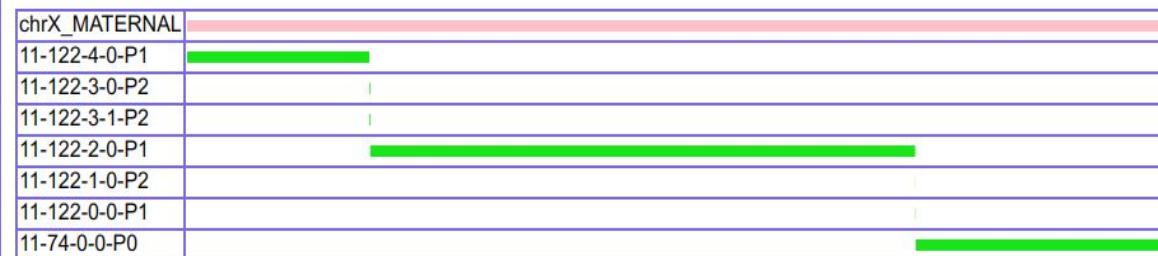
SHASTA v0.13.0 (HERRO)
September 2024

Previous Shasta release performed suboptimally on sex chromosomes

chrX

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



Alignments to chrX_MATERNAL

Shasta v0.14.0 (HERRO)

This reference segment is 154341397 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	
16-0-0-P0	

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
1121	154340868	154339747	16-0-0-P0	154332515	-	0	154332506	154332506	154328968	154343251	60	34	3504	10745	2.20e-07	2.27e-05	6.96e-05	0.9999	66.6	46.4	41.6	40.3

The new version of Shasta performs very well on the sex chromosomes

Alignments to chrY_PATERNAL

This reference segment is 62432599 bases long and has 1 alignments.

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

chrY_PATERNAL	
20-0-0-P0	

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
1180	62432599	62431419	20-0-0-P0	62430971	+	0	62429762	62429762	62429137	62432033	60	11	614	2271	1.76e-07	9.83e-06	3.64e-05	1.0000	67.5	50.1	44.4	43.3

VERKKO v2.2.1 (RAW + HERRO)

Alignments to chrX_MATERNAL

This reference segment is 154341397 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-0000581												

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	154341397	154341397	contig-0000581	154383726	-	2161	154380648	154378487	154330460	154389394	60	30	47997	10907	1.94e-07	3.11e-04	7.07e-05	0.9996	67.1	35.1	41.5	34.2

Alignments to chrY_PATERNAL

This reference segment is 62432599 bases long and has 1 alignments.

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

chrY_PATERNAL												
contig-0000533												

VERKKO performs very well on the sex chromosomes

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	62432599	62432598	contig-0000533	62434075	-	3294	62430586	62427292	62426725	62433153	60	12	555	5861	1.92e-07	8.89e-06	9.39e-05	0.9999	67.2	50.5	40.3	39.9

HIFIASM v0.24.0 (HERRO)

Alignments to chrX_MATERNAL

This reference segment is 154341397 bases long and has 12 alignments.

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



Alignments to chrY_PATERNAL

This reference segment is 62432599 bases long and has 5 alignments.

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



**Hifiasm assembles
some small extra
erroneous contigs**

**This happens across
all chromosomes**

**Hifiasm assembles chrY T2T
but
chrX is fragmented**

Sex chromosome assemblies compared to T2T HG002v1.1 Q100 Reference

chrX	SHASTA v0.14.0	VERKKO v2.2.1	Hifiasm v0.24.0
Mismatches	34	30	187
Insertions	3,504	47,997	3,812
Deletions	10,745	10,907	10,674

chrY	SHASTA v0.14.0	VERKKO v2.2.1	Hifiasm v0.24.0
Mismatches	11	12	22
Insertions	614	555	612
Deletions	2,271	5,861	2,197

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

Alignments to chr2_PATERNAL

This reference segment is 241873532 bases long and has 6 alignments.

Hifiasm generate some contigs with haplotype switches

Alignments to chr2_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr2_PATERNAL



Alignments to chr2_PATERNAL sorted by begin position in chr2_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
1707	4191164	4189457	h2tg000013l	242104013	+	6920	4192847	4185927	4161045	4209489	60	4850	20032	23562	1.16e-03	4.78e-03	5.62e-03	0.9885	29.4	23.2	22.5	19.4
4201976	87441212	83239236	h2tg000013l	242104013	+	4193015	87423095	83230080	83165666	83278648	60	25002	39412	48568	3.00e-04	4.73e-04	5.83e-04	0.9986	35.2	33.2	32.3	28.7
55002444	241873532	186871088	h1tg000016l	186870929	-	2833	186870929	186868096	186853258	186882864	60	3062	11776	14768	1.64e-05	6.30e-05	7.90e-05	0.9998	47.9	42.0	41.0	38.0
110155891	110191337	35446	h1tg000059l	35444	-	0	35444	35444	35443	35447	2	0	1	3	0.00e+00	2.82e-05	8.46e-05	0.9999	inf	45.5	40.7	39.5
115800688	116018000	217312	h1tg000110l	217311	+	0	217311	217311	217305	217317	51	1	5	6	4.60e-06	2.30e-05	2.76e-05	0.9999	53.4	46.4	45.6	42.6
164584571	164819359	234788	h1tg000127l	234774	-	0	234774	234774	234766	234792	7	4	4	18	1.70e-05	1.70e-05	7.67e-05	0.9999	47.7	47.7	41.2	39.6

*Mismatch Q phred score

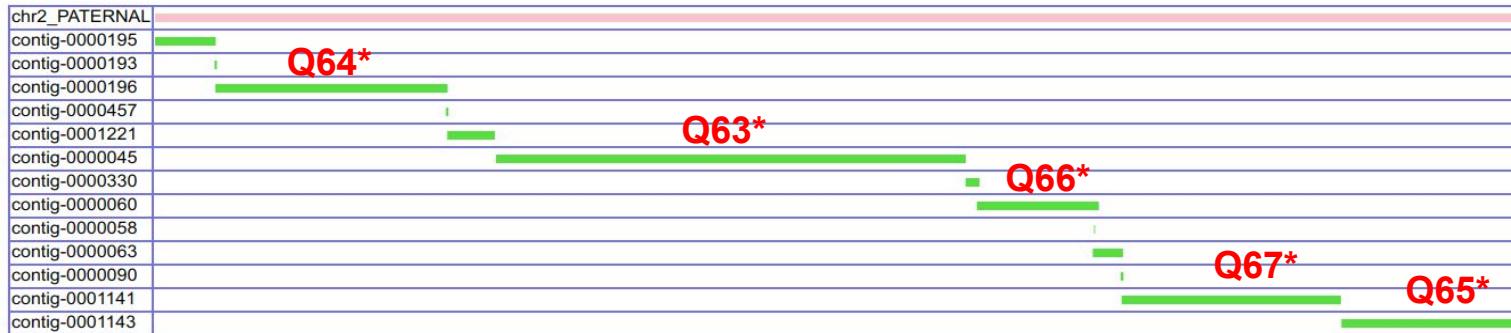
chr2

VERKKO v2.2.1 (RAW + HERRO)

Alignments to chr2_PATERNAL

This reference segment is 241873532 bases long and has 13 alignments.

Alignments to chr2_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr2_PATERNAL



Alignments to chr2_PATERNAL sorted by begin position in chr2_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			
0	10635989	10635989	contig-0000195	10638752	-	0	10635576	10635576	10635393	10636172	60	0	183	596	0.00e+00	1.72e-05	5.60e-05	0.9999	inf	47.6	42.5	41.4
10439810	10866968	427158	contig-0000193	426988	+	0	426988	426988	426971	427174	44	1	16	186	2.34e-06	3.75e-05	4.35e-04	0.9995	56.3	44.3	33.6	33.2
10576114	51316946	40740832	contig-0000196	40737955	+	0	40737926	40737926	40736878	40741866	60	14	1034	3940	3.44e-07	2.54e-05	9.67e-05	0.9999	64.6	46.0	40.1	39.1
51044508	51457409	412901	contig-0000457	412879	+	0	412879	412879	412875	412905	1	0	4	26	0.00e+00	9.69e-06	6.30e-05	0.9999	inf	50.1	42.0	41.4
51257575	59616005	8358430	contig-0001221	8357997	+	0	8357969	8357969	8357820	8358579	60	0	149	610	0.00e+00	1.78e-05	7.30e-05	0.9999	inf	47.5	41.4	40.4
59797978	142202069	82404091	contig-0000045	82399798	+	0	82399770	82399770	82397806	82406018	60	37	1927	6248	4.49e-07	2.34e-05	7.58e-05	0.9999	63.5	46.3	41.2	40.0
142180183	144625577	2445594	contig-0000330	2445246	+	0	2445215	2445215	2445195	2445413	18	1	19	198	4.09e-07	7.77e-06	8.10e-05	0.9999	63.9	51.1	40.9	40.5
144150135	165532536	21382401	contig-0000060	21381419	-	0	21381391	21381391	21380986	21382801	43	5	400	1410	2.34e-07	1.87e-05	6.59e-05	0.9999	66.3	47.3	41.8	40.7
164470990	169768064	5297074	contig-0000063	5296794	+	0	5296794	5296794	5296685	5297179	40	4	105	385	7.55e-07	1.98e-05	7.27e-05	0.9999	61.2	47.0	41.4	40.3
164709934	164895136	185202	contig-0000058	185189	+	0	185189	185189	185186	185204	5	1	2	15	5.40e-06	1.08e-05	8.10e-05	0.9999	52.7	49.7	40.9	40.1
169384109	169862342	478233	contig-0000090	478189	+	0	478189	478189	478173	478249	28	0	16	60	0.00e+00	3.35e-05	1.25e-04	0.9998	inf	44.8	39.0	38.0
169531840	208009667	38477827	contig-0001141	38475712	-	0	38475687	38475687	38474806	38478701	60	7	874	3014	1.82e-07	2.27e-05	7.83e-05	0.9999	67.4	46.4	41.1	39.9
208075315	241873532	33798217	contig-0001143	33799256	+	0	33796424	33796424	33795620	33799011	60	10	794	2587	2.96e-07	2.35e-05	7.65e-05	0.9999	65.3	46.3	41.2	40.0

*Mismatch Q phred score

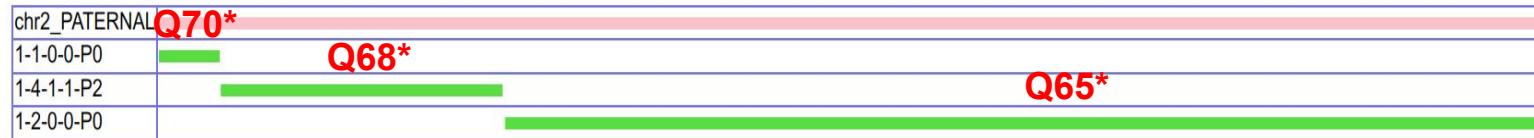
chr2

Shasta v0.14.0 (HERRO)

Alignments to chr2_PATERNAL

This reference segment is 241873532 bases long and has 3 alignments.

Alignments to chr2_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr2_PATERNAL



Alignments to chr2_PATERNAL sorted by begin position in chr2_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
115	10559759	10559644	1-1-0-0-P0	10559260	-	0	10559260	10559260	10559074	10559829	60	1	185	569	9.47e-08	1.75e-05	5.39e-05	0.9999	70.2	47.6	42.7	41.5
10739228	59657475	48918247	1-4-1-1-P2	48915090	-	0	48915090	48915090	48913859	48919470	60	8	1223	4380	1.64e-07	2.50e-05	8.95e-05	0.9999	67.9	46.0	40.5	39.4
60097720	241872728	181775008	1-2-0-0-P0	181765993	-	0	181765993	181765993	181761696	181779244	60	61	4236	13251	3.36e-07	2.33e-05	7.29e-05	0.9999	64.7	46.3	41.4	40.2

*Mismatch Q phred score

chr4

HIFIASM v0.24.0 (HERRO)

Alignments to chr4_PATERNAL

This reference segment is 192384017 bases long and has 2 alignments.

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

chr4_PATERNAL										
h1tg000070l	Q65*									
h1tg000006l										

Alignments to chr4_PATERNAL sorted by begin position in chr4_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
0	192384017	192384017	h1tg000006l	192384218	+	2575	192378118	192375543	192371660	192387840	60	60	3823	12297	3.12e-07	1.99e-05	6.39e-05	0.9999	65.1	47.0	41.9	40.8
22207666	22328045	120379	h1tg000070l	120373	-	0	120373	120373	120371	120379	3	2	0	6	1.66e-05	0.00e+00	4.98e-05	0.9999	47.8	inf	43.0	41.8

Alignments to chr4_MATERNAL

This reference segment is 191669995 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										
h2tg000001l	Q65*									
h1tg000011l										
h1tg000090l										

Alignments to chr4_MATERNAL sorted by begin position in chr4_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	191669995	191669995	h2tg000001l	191668051	+	3420	191665178	191661758	191657757	191673936	60	60	3941	12178	3.13e-07	2.06e-05	6.35e-05	0.9999	65.0	46.9	42.0	40.7
148567882	149244539	676657	h1tg000011l	676602	+	0	676602	676602	676587	676672	24	0	15	70	0.00e+00	2.22e-05	1.03e-04	0.9999	inf	46.5	39.9	39.0
163890353	164218894	328541	h1tg000090l	328529	+	0	328529	328529	328520	328550	60	0	9	21	0.00e+00	2.74e-05	6.39e-05	0.9999	inf	45.6	41.9	40.4

*Mismatch Q phred score

Alignments to chr4_PATERNAL

This reference segment is 192384017 bases long and has 3 alignments.

Alignments to chr4_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr4_PA



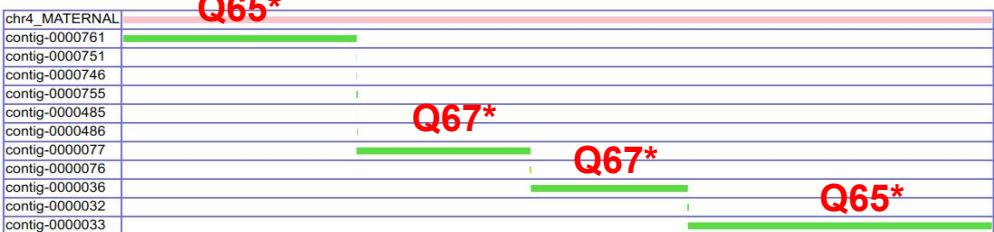
Alignments to chr4_PATERNAL sorted by begin position in chr4_PATERNAL

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	
0	90699900	90699900	contig-0000853	90697874	+	2563	90697846	90695283	90693553	90701615	60	15	1715	6332	1.65e-07	1.89e-05	6.98e-05	0.9999	67.8	47.2	41.6	40.5
90481621	125396326	34914705	contig-0000035	34913157	-	0	34913128	34913128	34912536	34915281	60	16	576	2153	4.58e-07	1.65e-05	6.17e-05	0.9999	63.4	47.8	42.1	41.0
125082518	192384017	67301499	contig-0000034	67304473	+	0	67298374	67298374	67297073	67302790	60	10	1291	4416	1.49e-07	1.92e-05	6.56e-05	0.9999	68.3	47.2	41.8	40.7

Alignments to chr4_MATERNAL

This reference segment is 191669995 bases long and has 11 alignments.

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



Alignments to chr4_MATERNAL sorted by begin position in chr4_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	51547159	51547159	contig-0000761	51548059	-	0	51544640	51544640	51543555	51548228	60	16	1069	3588	3.10e-07	2.07e-05	6.96e-05	0.9999	65.1	46.8	41.6	40.4
51388275	51467402	79127	contig-0000751	79126	+	0	79126	79126	79126	79127	4	0	0	1	0.00e+00	0.00e+00	1.26e-05	1.0000	inf	inf	49.0	49.0
51242305	51761777	337472	contig-0000755	337472	+	0	337472	337472	337470	337472	12	2	0	0	5.93e-06	0.00e+00	0.00e+00	1.0000	52.3	inf	inf	52.3
51443084	51536108	93024	contig-0000746	93054	-	0	93024	93024	93022	93024	6	2	0	0	2.15e-05	0.00e+00	0.00e+00	1.0000	46.7	inf	inf	46.7
51462235	89865341	38403106	contig-000077	38401314	-	0	38401283	38401283	38400593	38403789	60	7	683	2506	1.82e-07	1.78e-05	6.53e-05	0.9999	67.4	47.5	41.9	40.8
51608035	51651563	43528	contig-0000485	43560	+	0	43528	43528	43524	43528	11	4	0	0	9.19e-05	0.00e+00	0.00e+00	0.9999	40.4	inf	inf	40.4
51622979	51812568	189589	contig-0000486	189619	+	0	189589	189589	189584	189589	44	5	0	0	2.64e-05	0.00e+00	0.00e+00	1.0000	45.8	inf	inf	45.8
89592008	90051188	459180	contig-0000076	459162	-	0	459162	459162	459037	459240	1	65	60	78	1.42e-04	1.31e-04	1.70e-04	0.9996	38.5	38.8	37.7	33.5
89926906	124568959	34642053	contig-0000036	34640474	-	0	34640446	34640446	34639893	34642600	60	6	547	2154	1.73e-07	1.58e-05	6.22e-05	0.9999	67.6	48.0	42.1	41.1
124447078	124778698	331620	contig-0000032	331593	-	0	331593	331593	331587	331625	5	1	5	32	3.02e-06	1.51e-05	9.65e-05	0.9999	55.2	48.2	40.2	39.4
124570607	191669995	67099388	contig-0000033	67099085	+	11	67096213	67096202	67094870	67100701	60	19	1313	4499	2.83e-07	1.96e-05	6.70e-05	0.9999	65.5	47.1	41.7	40.6

*Mismatch Q phred score

VERKKO v2.2.1 (RAW + HERRO)

Without extra HiC/PoreC data,
Verkko makes more
conservative choices regarding
phasing, leading to lower N50
but producing contigs with
good accuracy

chr4

Shasta v0.14.0 (HERRO)

Alignments to chr4_PATERNAL

This reference segment is 192384017 bases long and has 1 alignments.

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

chr4_PATERNAL	Q67*						
4-1-0-0-P0							

Alignments to chr4_PATERNAL sorted by begin position in chr4_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			
0	192384017	192384017	4-1-0-0-P0	192378958	+	1687	192376937	192375250	192371475	192387752	60	40	3735	12502	2.08e-07	1.94e-05	6.50e-05	0.9999	66.8	47.1	41.9	40.7

Alignments to chr4_MATERNAL

This reference segment is 191669995 bases long and has 1 alignments.

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

chr4_MATERNAL	Q66*						
4-0-0-0-P0							

Alignments to chr4_MATERNAL sorted by begin position in chr4_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	191669995	191669995	4-0-0-0-P0	191663778	+	235	191661656	191661421	191657577	191673794	60	45	3799	12373	2.35e-07	1.98e-05	6.46e-05	0.9999	66.3	47.0	41.9	40.7

Shasta v0.14.0 manages to assemble both haplotypes T2T while maintaining its good accuracy performance

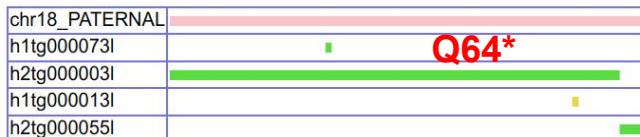
*Mismatch Q phred score

Alignments to chr18_PATERNAL

HIFIASM v0.24.0 (HERRO)

This reference segment is 80778407 bases long and has 4 alignments.

Alignments to chr18_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr18_PATERNAL



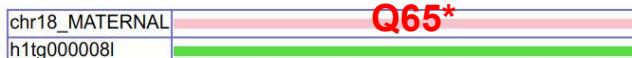
Alignments to chr18_PATERNAL sorted by begin position in chr18_PATERNAL

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
0	75851635	75851635	h2tg000003I	75852217	-	0	75848260	75848260	75846666	75853200	60	29	1565	4940	3.82e-07	2.06e-05	6.51e-05	0.9999	64.2	46.9	41.9	40.6
26268283	27290771	1022488	h1tg000073I	1022410	+	0	1022410	1022410	1022384	1022513	35	1	25	103	9.78e-07	2.45e-05	1.01e-04	0.9999	60.1	46.1	40.0	39.0
67862430	68928450	1066020	h1tg000013I	1066109	-	0	1066109	1066109	1065396	1066344	15	389	324	235	3.65e-04	3.04e-04	2.20e-04	0.9991	34.4	35.2	36.6	30.5
75865229	80778407	4913178	h2tg000055I	4915572	+	0	4913039	4913039	4912934	4913277	60	6	99	238	1.22e-06	2.01e-05	4.84e-05	0.9999	59.1	47.0	43.1	41.6

Alignments to chr18_MATERNAL

This reference segment is 78908764 bases long and has 1 alignments.

Alignments to chr18_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr18_MATERNAL



Alignments to chr18_MATERNAL sorted by begin position in chr18_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	78868707	78868707	h1tg000008I	78869146	-	0	78865430	78865430	78863665	78870451	60	21	1744	5021	2.66e-07	2.21e-05	6.37e-05	0.9999	65.7	46.6	42.0	40.7

*Mismatch Q phred score

VERKKO v2.2.1 (RAW + HERRO)

Alignments to chr18_PATERNAL

This reference segment is 80778407 bases long and has 1 alignments.

Alignments to chr18_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr18_PATERNAL

chr18_PATERNAL	Q64*						
contig-0000438							

Alignments to chr18_PATERNAL sorted by begin position in chr18_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			
0	80778407	80778407	contig-0000438	80780965	+	3957	80778433	80774476	80772938	80779914	60	31	1507	5438	3.84e-07	1.87e-05	6.73e-05	0.9999	64.2	47.3	41.7	40.6

Alignments to chr18_MATERNAL

This reference segment is 78908764 bases long and has 1 alignments.

Alignments to chr18_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr18_MATERNAL

chr18_MATERNAL	Q59*						
contig-0000676							

Alignments to chr18_MATERNAL sorted by begin position in chr18_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	78908764	78908764	contig-0000676	78913230	-	2503	78909515	78907012	78903392	78912292	60	92	3528	5280	1.17e-06	4.47e-05	6.69e-05	0.9999	59.3	43.5	41.7	39.5

*Mismatch Q phred score

Shasta v0.14.0 (HERRO)

Alignments to chr18_PATERNAL

This reference segment is 80778407 bases long and has 1 alignments.

Alignments to chr18_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr18_PATERNAL

chr18_PATERNAL	Q64*	
15-1-0-0-P0		

Alignments to chr18_PATERNAL sorted by begin position in chr18_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
266	80778407	80778141	15-1-0-0-P0	80776254	+	0	80774486	80774486	80772830	80779765	60	32	1624	5279	3.96e-07	2.01e-05	6.54e-05	0.9999	64.0	47.0	41.8	40.7

Alignments to chr18_MATERNAL

This reference segment is 78908764 bases long and has 1 alignments.

Alignments to chr18_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr18_MATERNAL

chr18_MATERNAL	Q66*	
15-0-0-0-P0		

Alignments to chr18_MATERNAL sorted by begin position in chr18_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
611	78907748	78907137	15-0-0-0-P0	78903568	+	0	78903568	78903568	78901967	78908720	60	18	1583	5152	2.28e-07	2.01e-05	6.53e-05	0.9999	66.4	47.0	41.9	40.7

*Mismatch Q phred score

Compared to Verkko, Shasta

HERRO has:

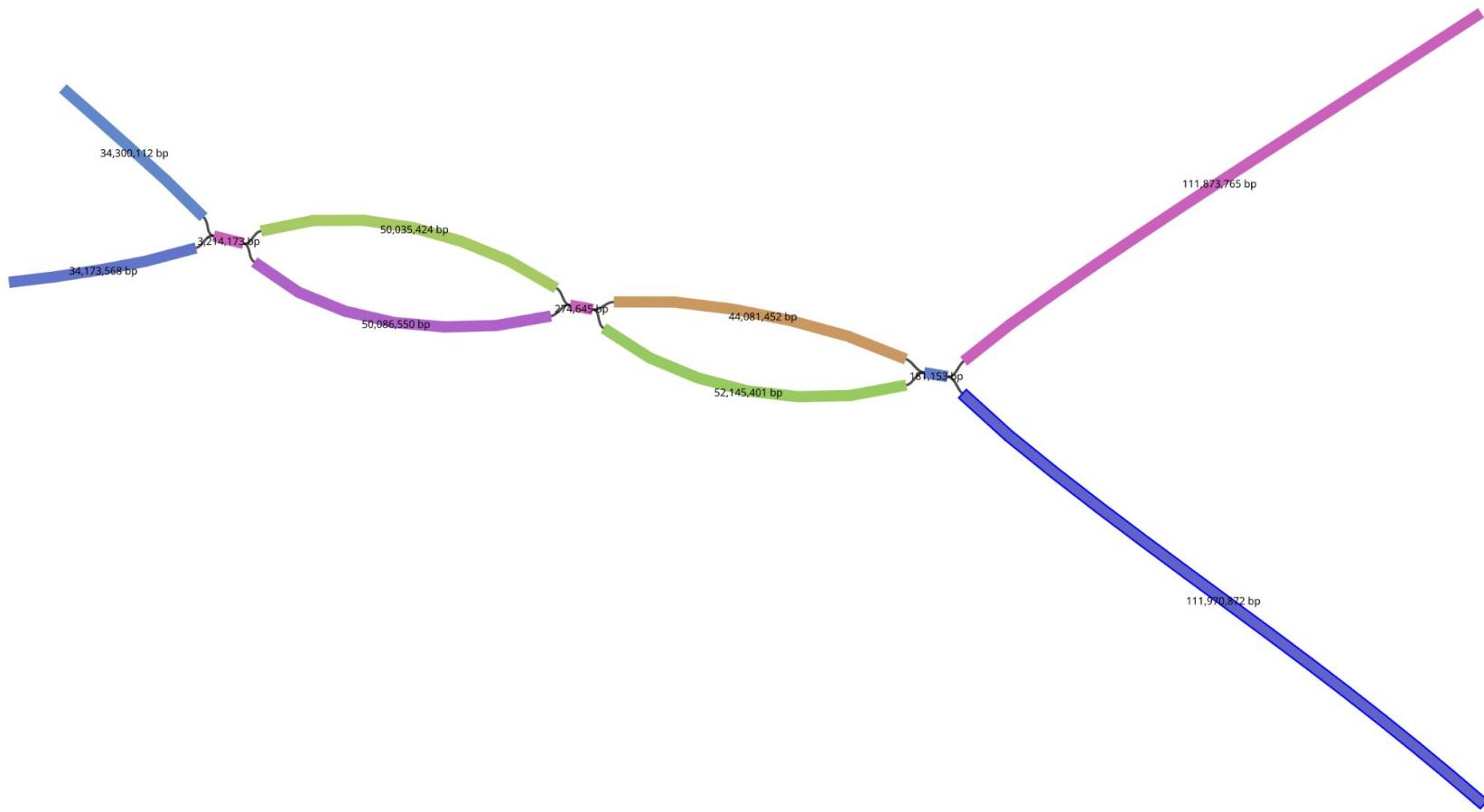
- 5 times fewer mismatches
 - Half the insertions
 - Fewer deletions

**Shasta v0.14.0 achieves near T2T assembly on
chrX, chrY, chr4, chr16, chr18
while being consistently highly accurate**

Let's see a few other examples

chr1

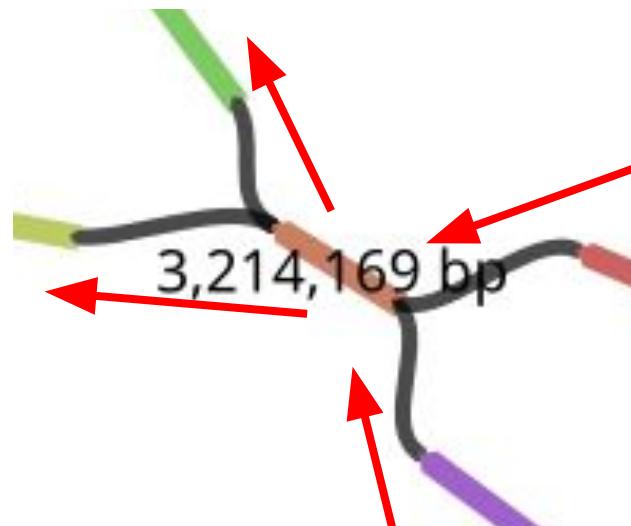
Shasta v0.14.0 (HERRO)



Unphased region

Possible reasons:

- The reads are not long enough
- Heterozygosity is too low in that region
- Coverage is not high enough



Note: We can see several cases with coverage loss on only one haplotype

Alignments to chr1_PATERNAL

This reference segment is 252060741 bases long and has 5 alignments.

Mismatches: Q65 ~1 mismatch per 4 million bases

Alignments to chr1_PATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr1_PATERNAL

chr1_PATERNAL												
2-2-0-0-P0												
2-4-0-0-P1												
2-4-1-1-P2												
2-4-3-1-P2												
2-1-0-0-P0												



Alignments to chr1_PATERNAL sorted by begin position in chr1_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
804	111978057	111977253	2-2-0-0-P0	111970872	+	0	111970872	111970872	111967372	111980729	60	24	3476	9857	2.14e-07	3.10e-05	8.80e-05	0.9999	66.7	45.1	40.6	39.2
111978055	112159204	181149	2-4-0-0-P1	181153	+	0	181153	181153	181146	181156	1	0	7	3	0.00e+00	3.86e-05	1.66e-05	0.9999	inf	44.1	47.8	42.6
112159202	164306379	52147177	2-4-1-1-P2	52145401	+	0	52145401	52145401	52144512	52148052	60	14	875	2651	2.68e-07	1.68e-05	5.08e-05	0.9999	65.7	47.8	42.9	41.7
164632548	214670601	50038053	2-4-3-1-P2	50035424	+	0	50035424	50035424	50034253	50039213	60	11	1160	3789	2.20e-07	2.32e-05	7.57e-05	0.9999	66.6	46.3	41.2	40.0
217884940	252060288	34175348	2-1-0-0-P0	34173568	+	0	34173553	34173553	34172403	34176476	60	22	1128	2923	6.44e-07	3.30e-05	8.55e-05	0.9999	61.9	44.8	40.7	39.2

Alignments to chr1_MATERNAL

This reference segment is 244022067 bases long and has 6 alignments.

Mismatches: Q67 ~1 mismatch per 5 million bases

Alignments to chr1_MATERNAL sorted by assembled segment, and for each assembled segment by begin position in chr1_MATERNAL

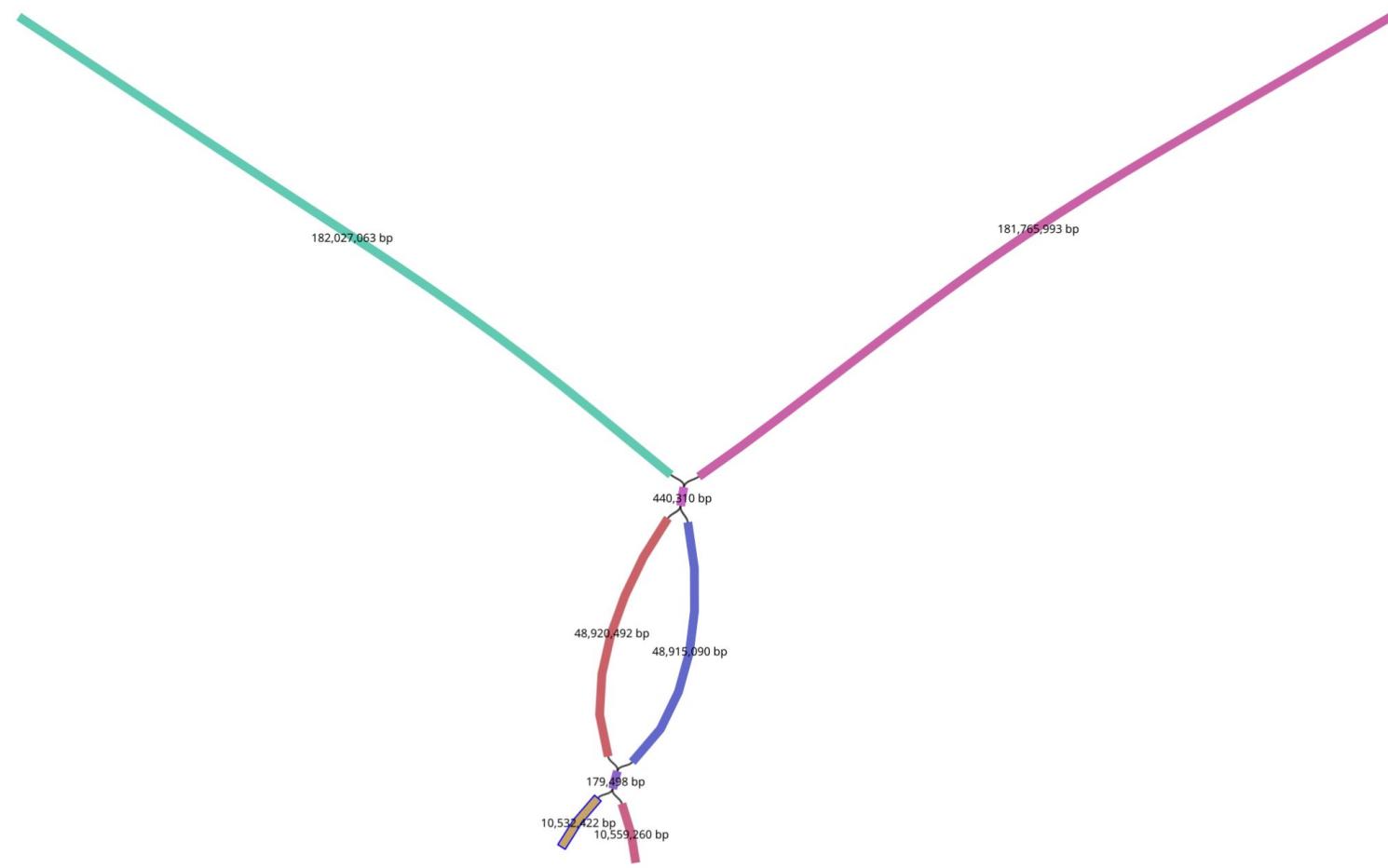
chr1_MATERNAL												
2-3-0-0-P0												
2-4-1-0-P2												
2-4-2-0-P1												
2-4-3-0-P2												
2-4-4-0-P1												
2-0-0-0-P0												



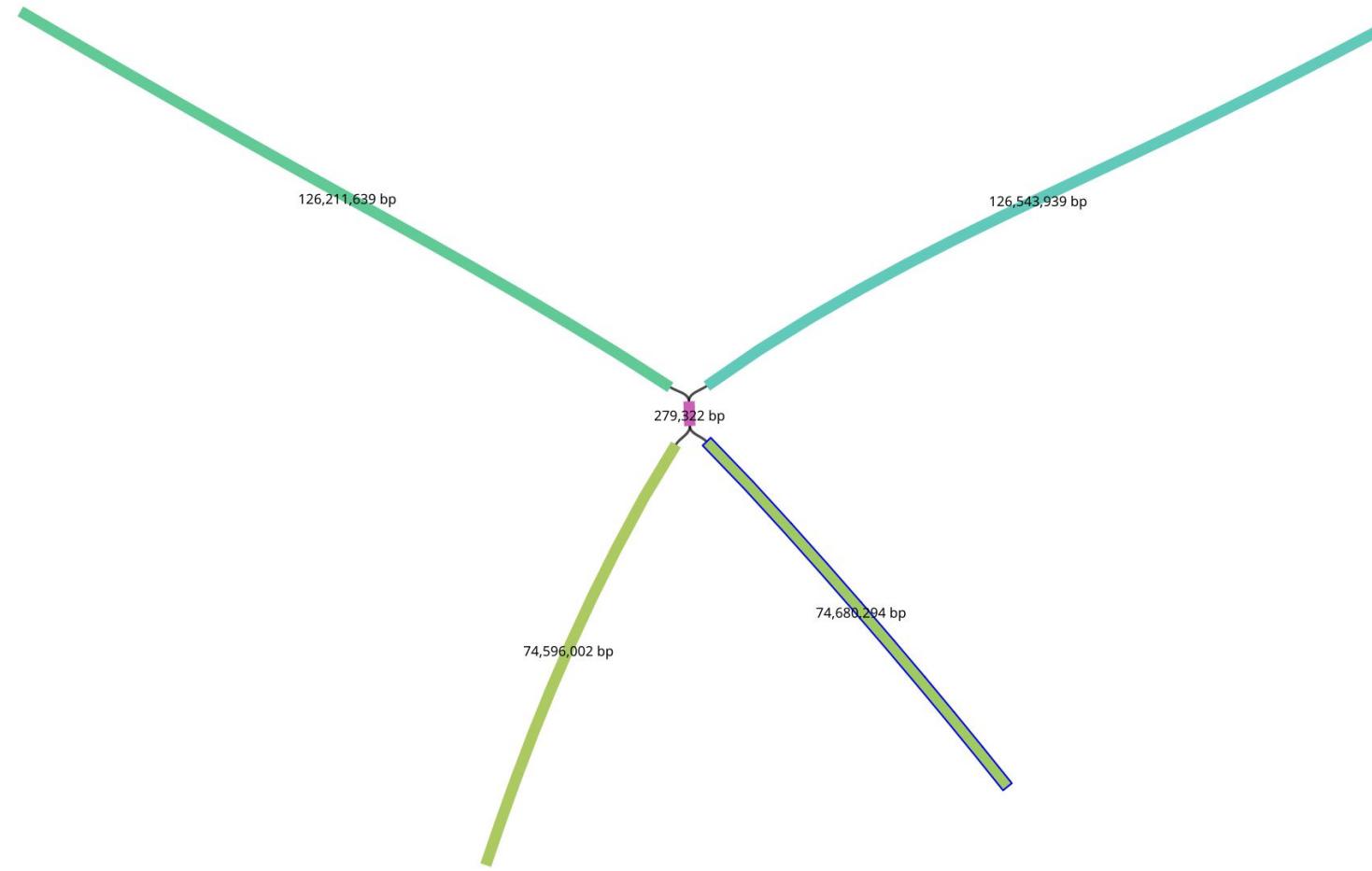
Alignments to chr1_MATERNAL sorted by begin position in chr1_MATERNAL

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
2	111879365	111879363	2-3-0-0-P0	111873765	+	0	111873765	111873765	111869581	111883508	60	39	4145	9743	3.49e-07	3.70e-05	8.71e-05	0.9999	64.6	44.3	40.6	39.0
112060520	156143652	440813132	2-4-1-0-P2	44081452	+	0	44081452	44081452	44080435	44084141	60	8	1009	2689	1.81e-07	2.29e-05	6.10e-05	0.9999	67.4	46.4	42.1	40.8
156143623	156418286	274663	2-4-2-0-P1	274645	+	0	274645	274645	274635	274673	36	0	10	28	0.00e+00	3.64e-05	1.02e-04	0.9999	inf	44.4	39.9	38.6
156418281	206507388	50089107	2-4-3-0-P2	50086550	+	0	50086550	50086550	50085350	50090294	60	13	1187	3744	2.60e-07	2.37e-05	7.47e-05	0.9999	65.9	46.3	41.3	40.1
206507364	209721747	3214383	2-4-4-0-P1	3214173	+	0	3214173	3214173	3214103	3214452	1	1	69	279	3.11e-07	2.15e-05	8.68e-05	0.9999	65.1	46.7	40.6	39.6
209721740	244022067	34300327	2-0-0-0-P0	34300112	+	0	34298550	34298550	34297486	34301382	60	9	1055	2832	2.62e-07	3.08e-05	8.26e-05	0.9999	65.8	45.1	40.8	39.4

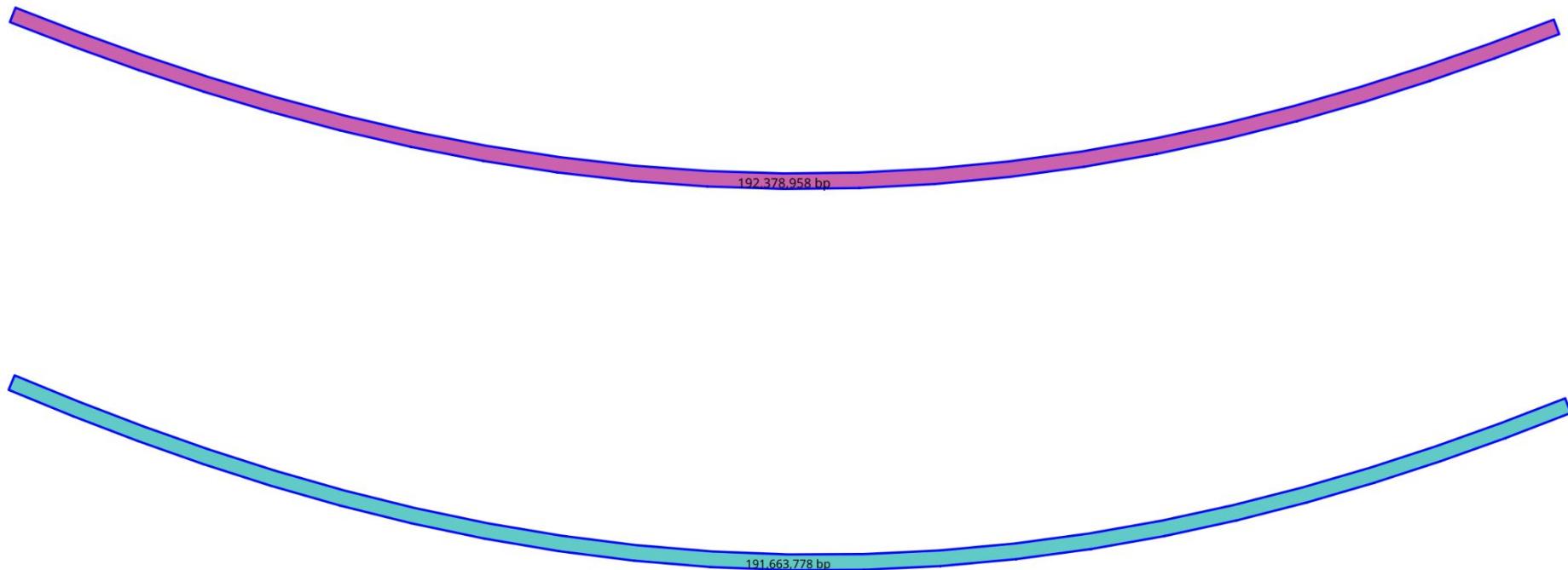
chr2



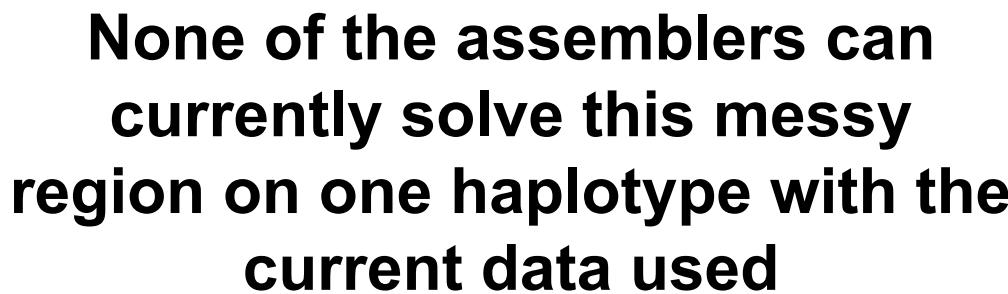
chr3



chr4

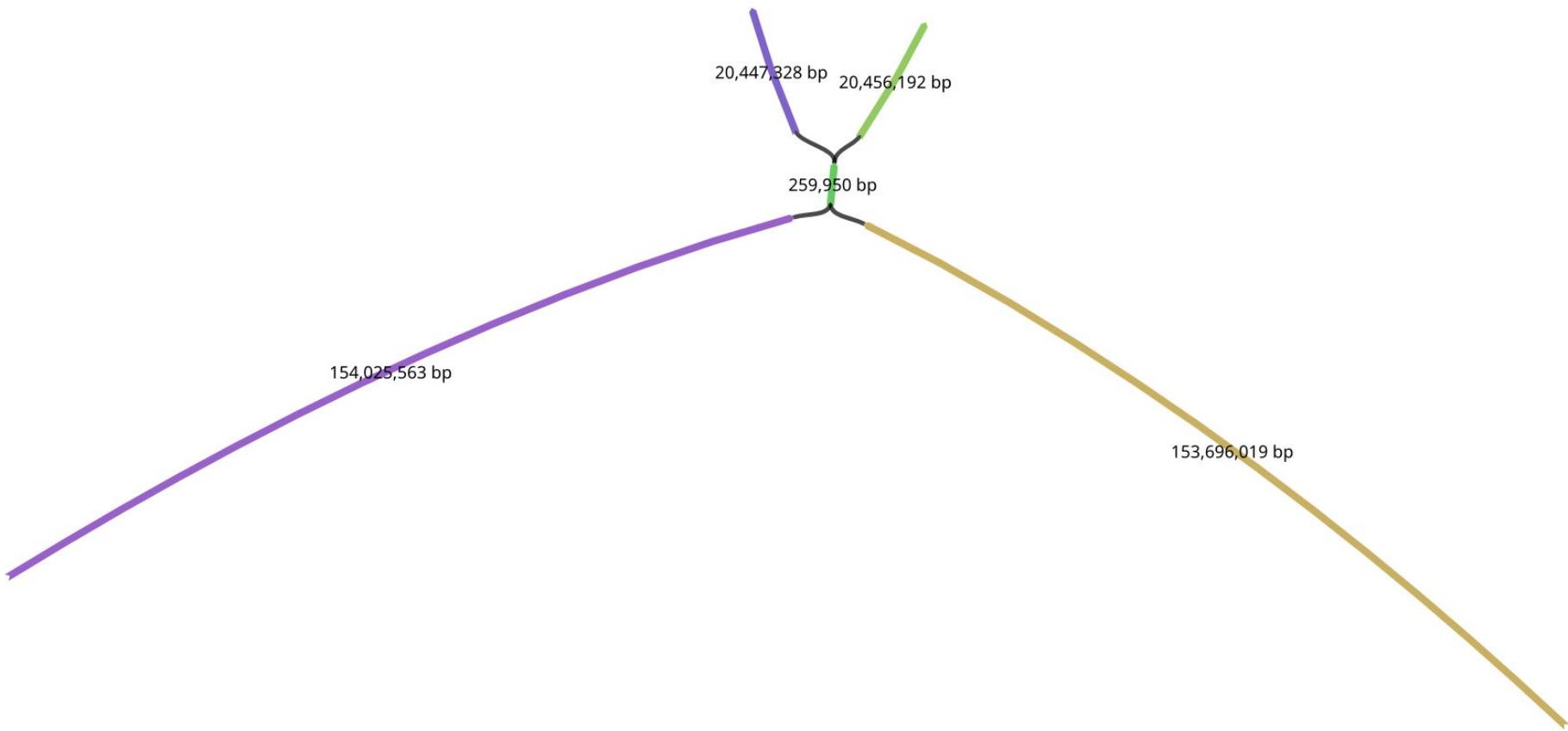


chr5



None of the assemblers can currently solve this messy region on one haplotype with the current data used

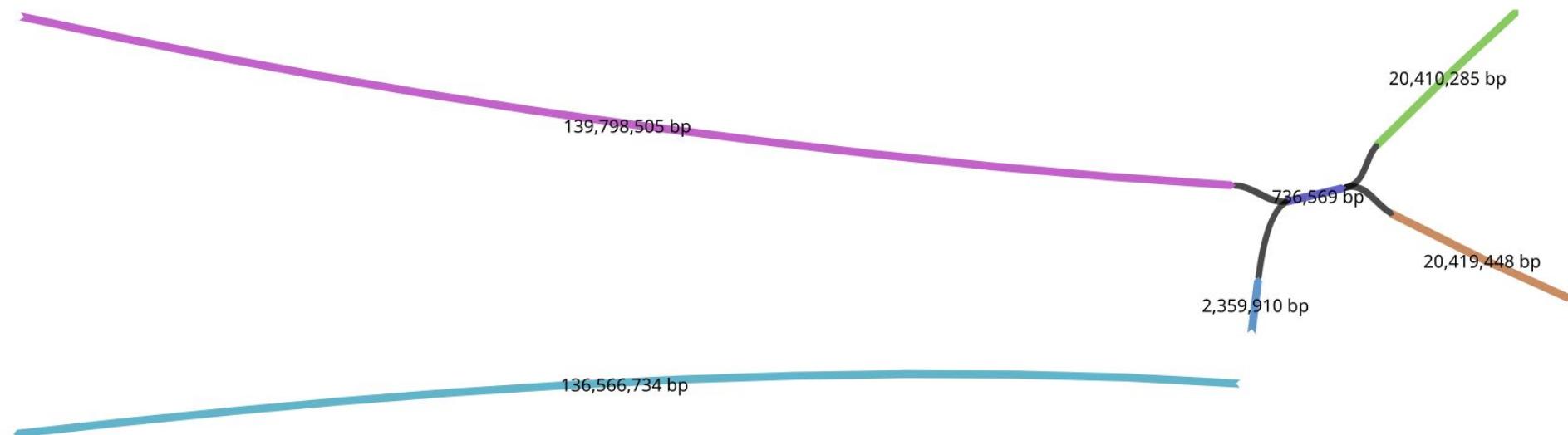
chr6



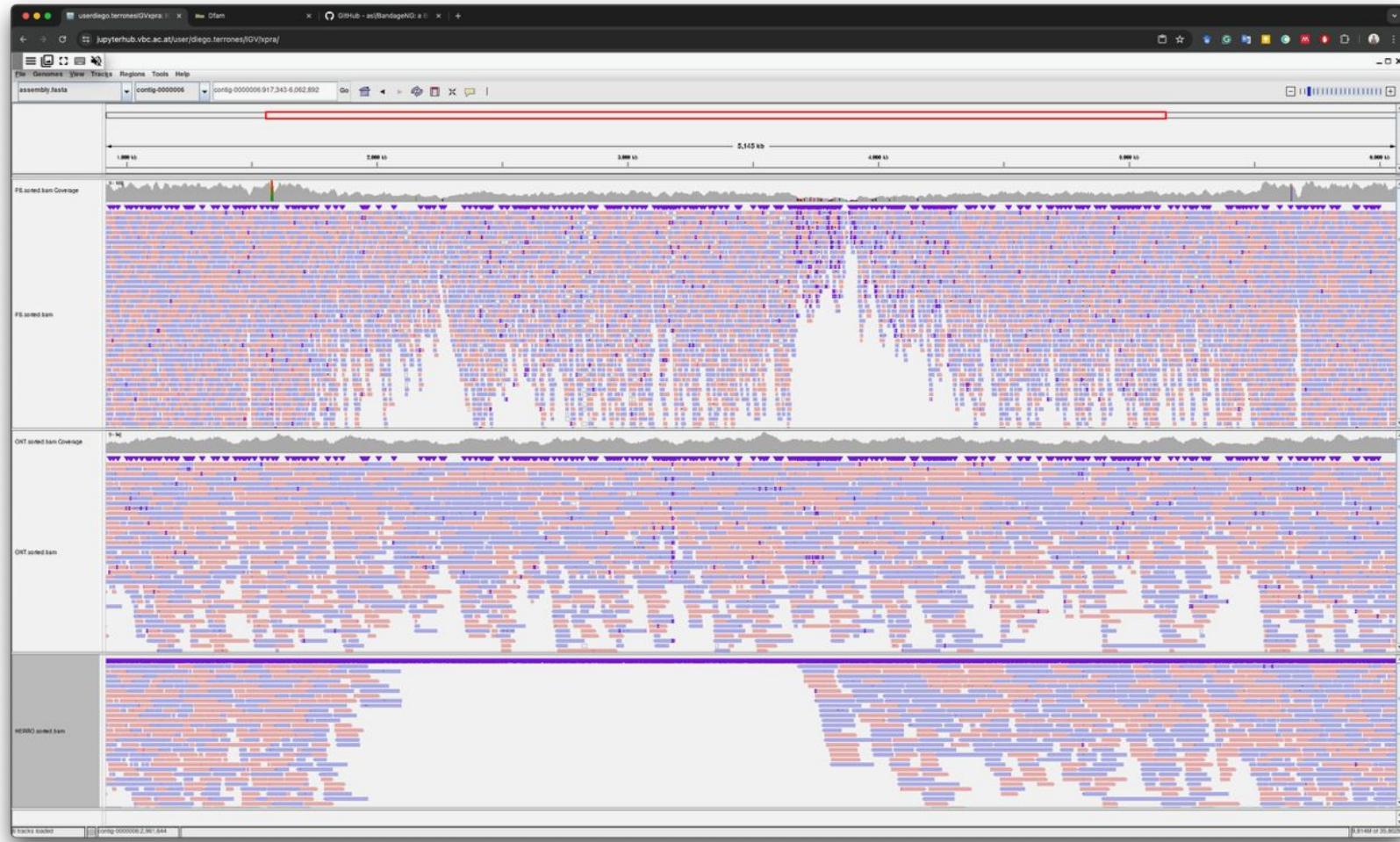
Known issues that need fixing

Occasionally we get a break on one haplotype due to coverage loss

We are working on even more robust read graph methods while maintaining the same accuracy and speed



HERRO discards reads in highly repetitive regions => minimap2 issue



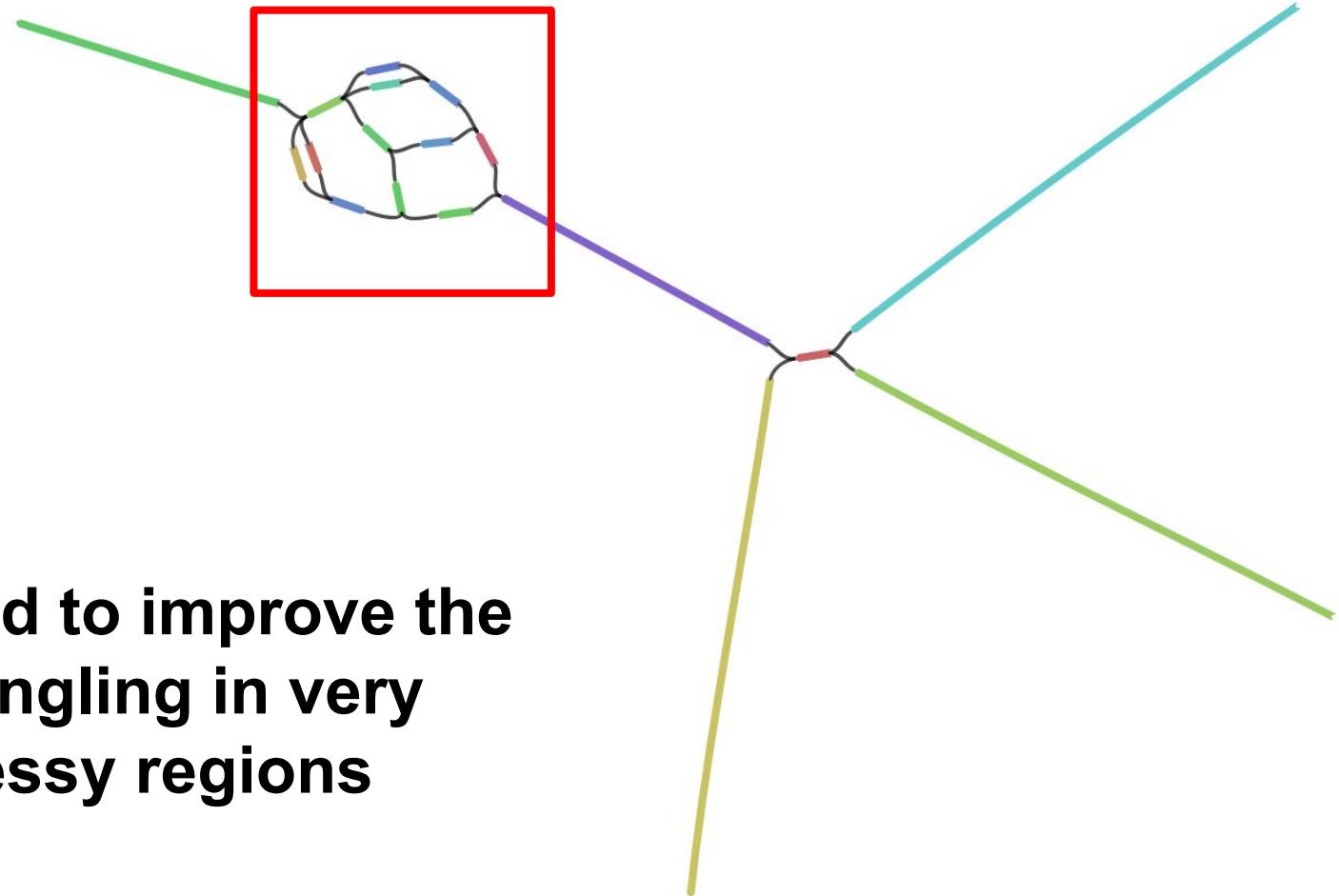
ONT
HERRO

What about telomeres?

We are missing in most cases the last
~500-1500 telomeric bases

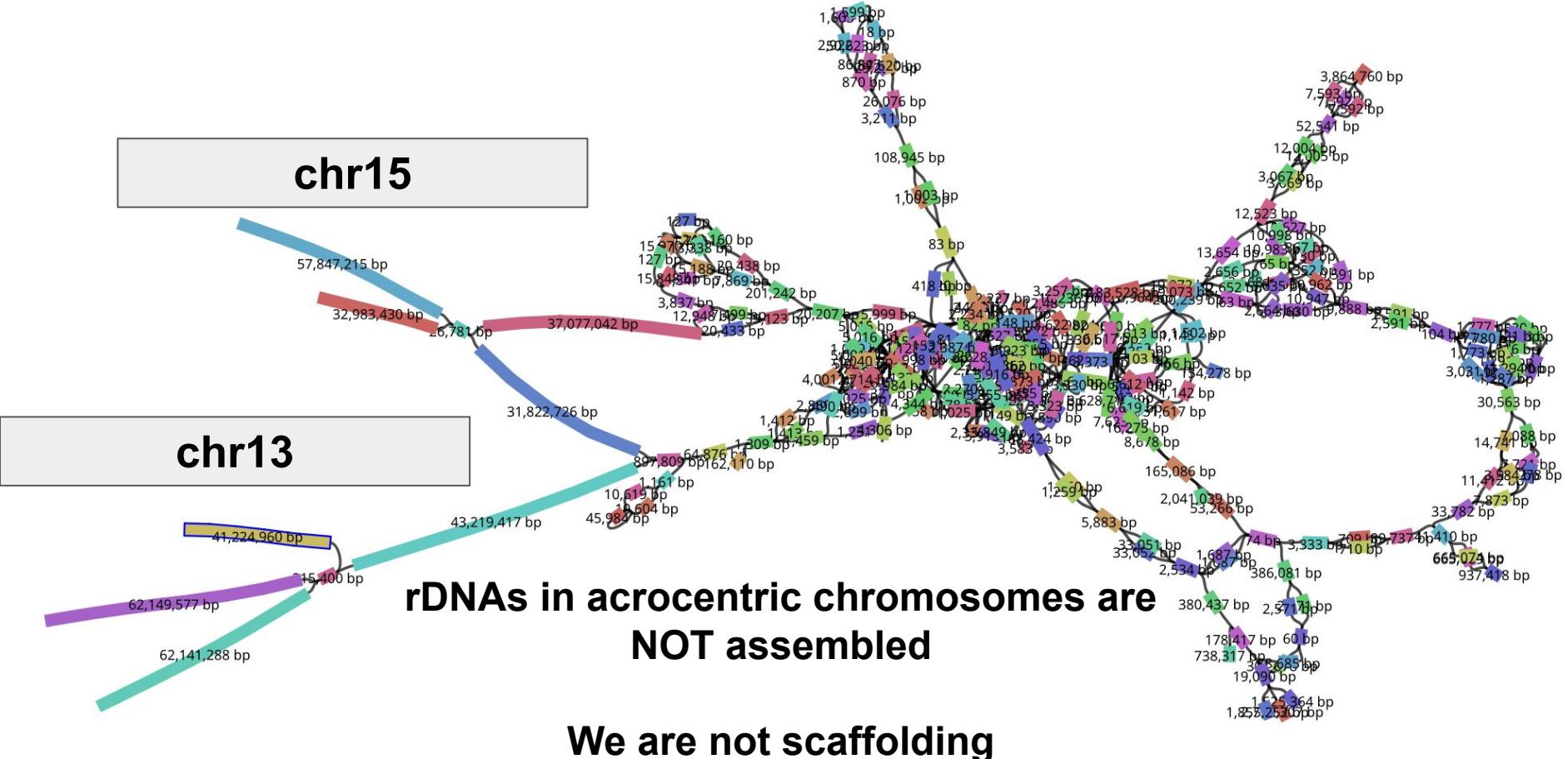
This will be fixed in future releases

chr5



We need to improve the
detangling in very
messy regions

What about acrocentric chromosomes?



Looking Ahead... Complete Assembly

Prototype T2T Shasta protocol:

- **2 ONT UL flow cells** using mod prep achieve **~40x coverage reads >100KB**
- **1 Hi-C/Pore-C flow cell**
- Estimated cost around **~\$5k**
- New Shasta “mode 3” - **3 hours on 1 compute node**
- **137.5 MB NG50 - 33 / 46 validated T2T chromosomes (with Hi-C)**

We have a path to achieve the same results using only 1FC over the next three years for a total budget of \$1,000

Acknowledgements

UNIVERSITY OF CALIFORNIA
SANTA CRUZ Genomics
Institute



Paolo Carnevali



Miten Jain



Ryan Lorig-Roach



Mikhail Kolmogorov



La science pour la santé
From science to health



Jean Monlong



Karen Miga



Julian Lucas

<https://cglgenomics.ucsc.edu/>



Adam Novak



Jordan Eizenga



David Haussler



Melissa Meredith



Parsa Eskander



Glenn Hickey



Mark Diekhans



Xian Chang



Jouni Siren



Brandy Baird
Joshua Gardner
Sara O'Rourke