Telomere-to-Telomere Assembly of the SHRSP/BbbUtx (SHR-A3) Rat

Kai Li
Graduate Student
Department of Veterinary Science
Martin-Gatton College of Agriculture, Food, and Environment
University of Kentucky

Introduction:

Genome Assembly

Goal: produce **contiguous and complete** genomes, covering complex and repetitive regions.

Challenges: Traditional methods struggle with **complex/repetitive regions** (Telomeres, Centromeres, rDNA arrays)

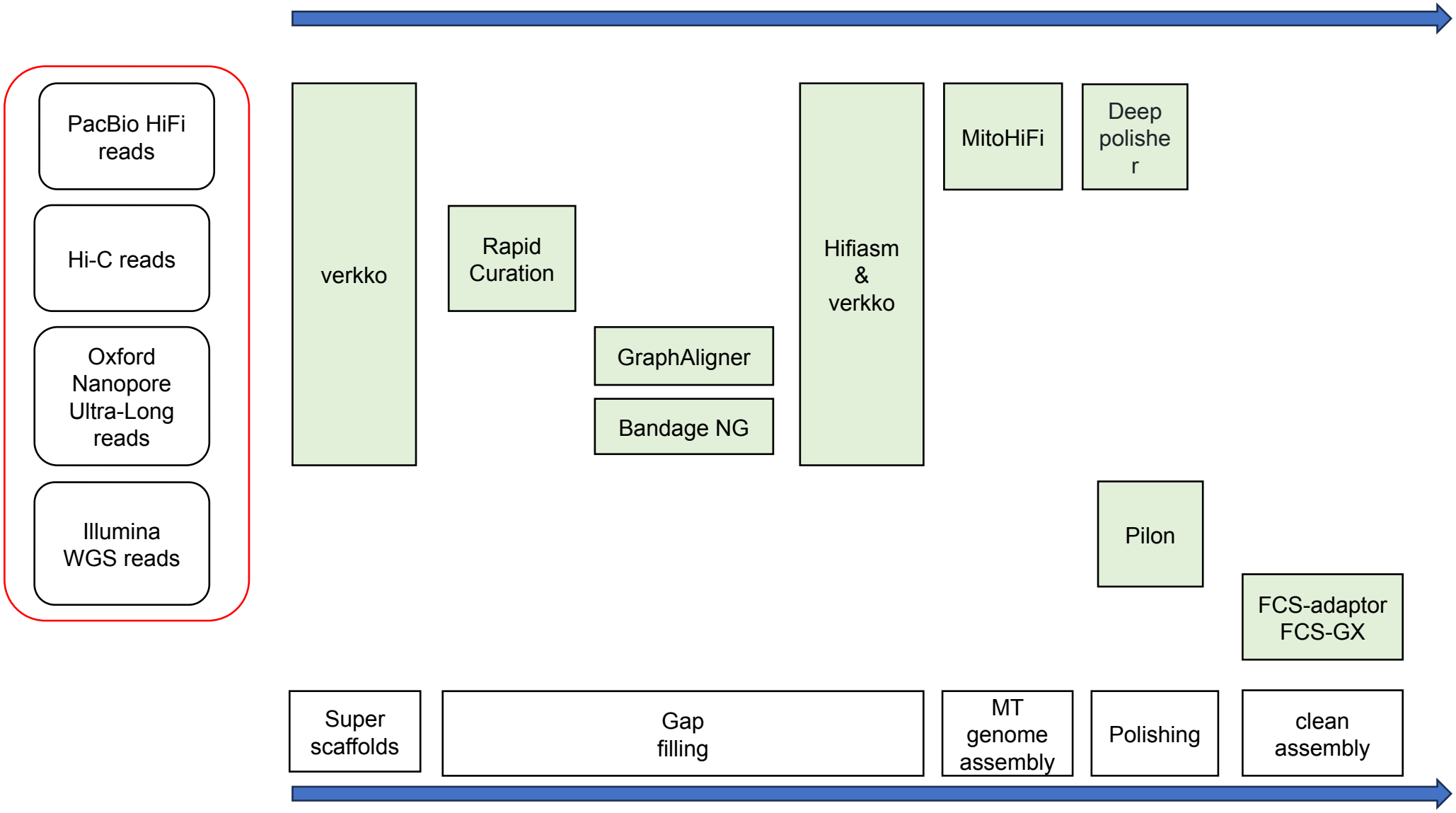
Telomere-to-Telomere (T2T) Assembly

- Take advantage of Oxford Nanopore ultra long sequencing to overcome challenges with complex/repetitive regions.
- T2T assembly = full chromosome assembly from one telomere to the other.

Data Summary:

Data Type	Reads (M)	Total (Gb)	Coverage
PacBio HiFi	7.67	96.52	~34X
Hi-C	1617.8	242.65	~87X
Oxford Nanopore Ultra-Long	8.49	241.5	~32X
Illumina Short Read WGS	1306.7	128.5	~46X

Assembly Method:

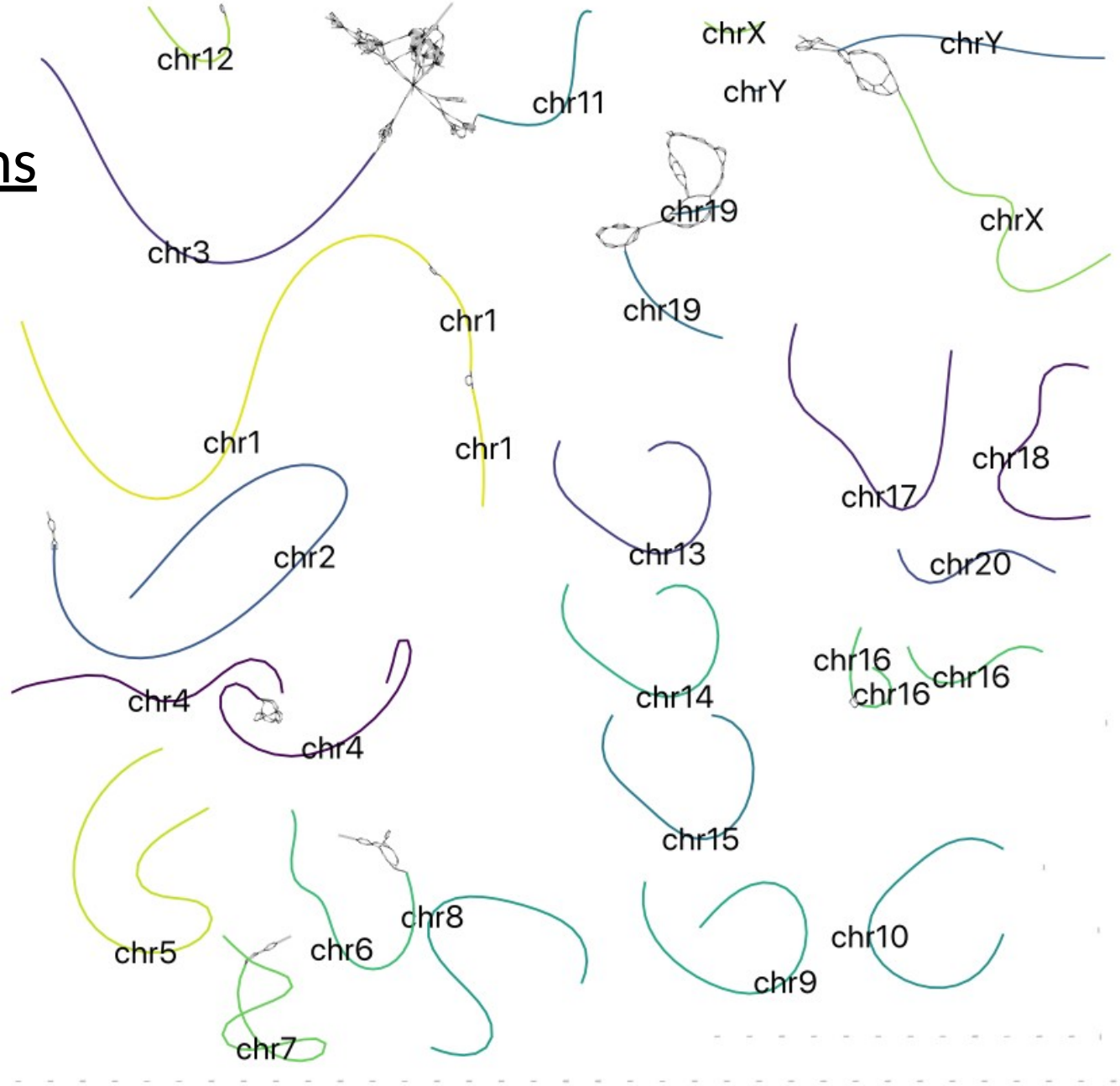


Visualizing assembly graphs

Bandage NG

Visualizing assembly graphs

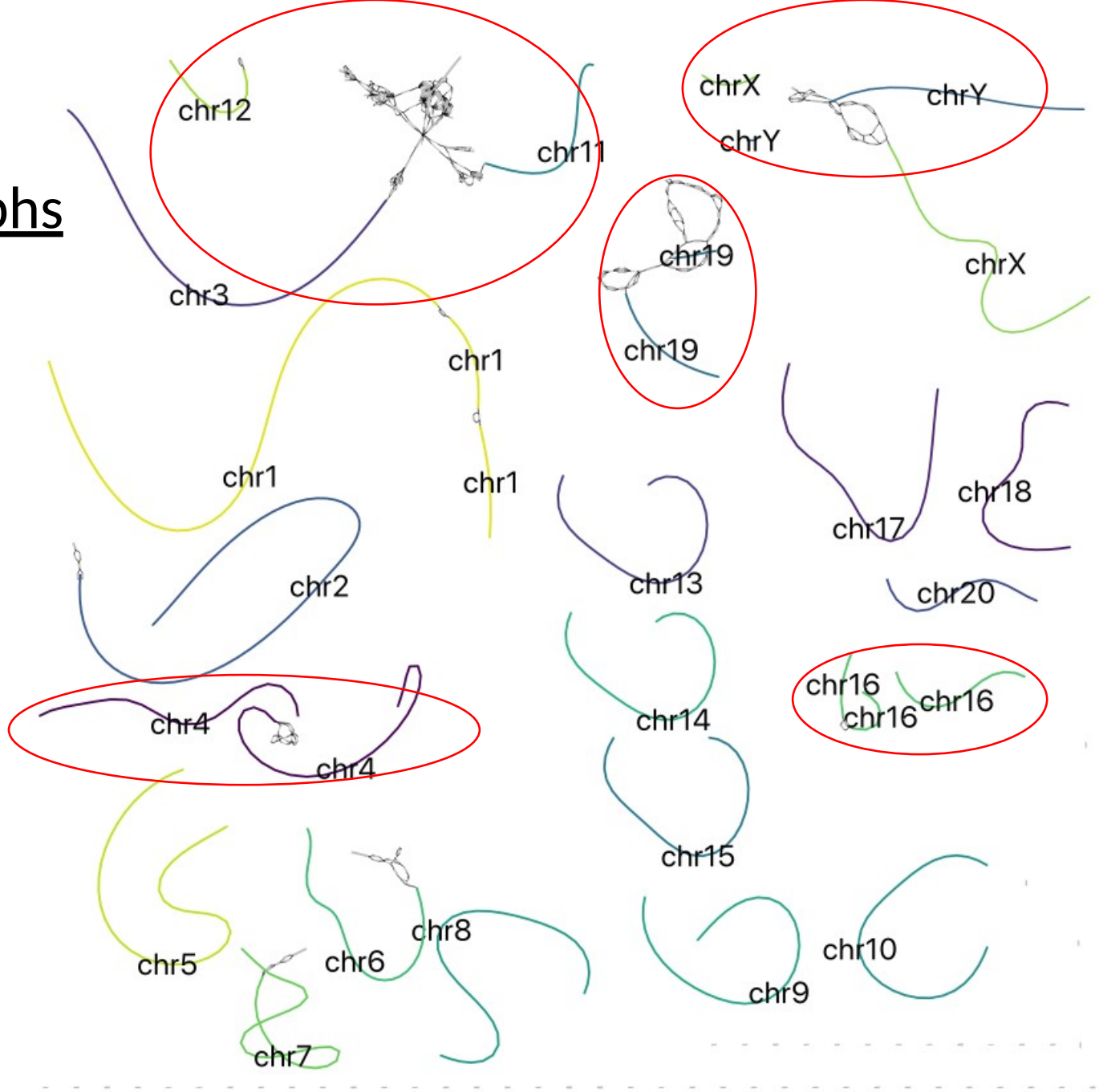
Bandage NG



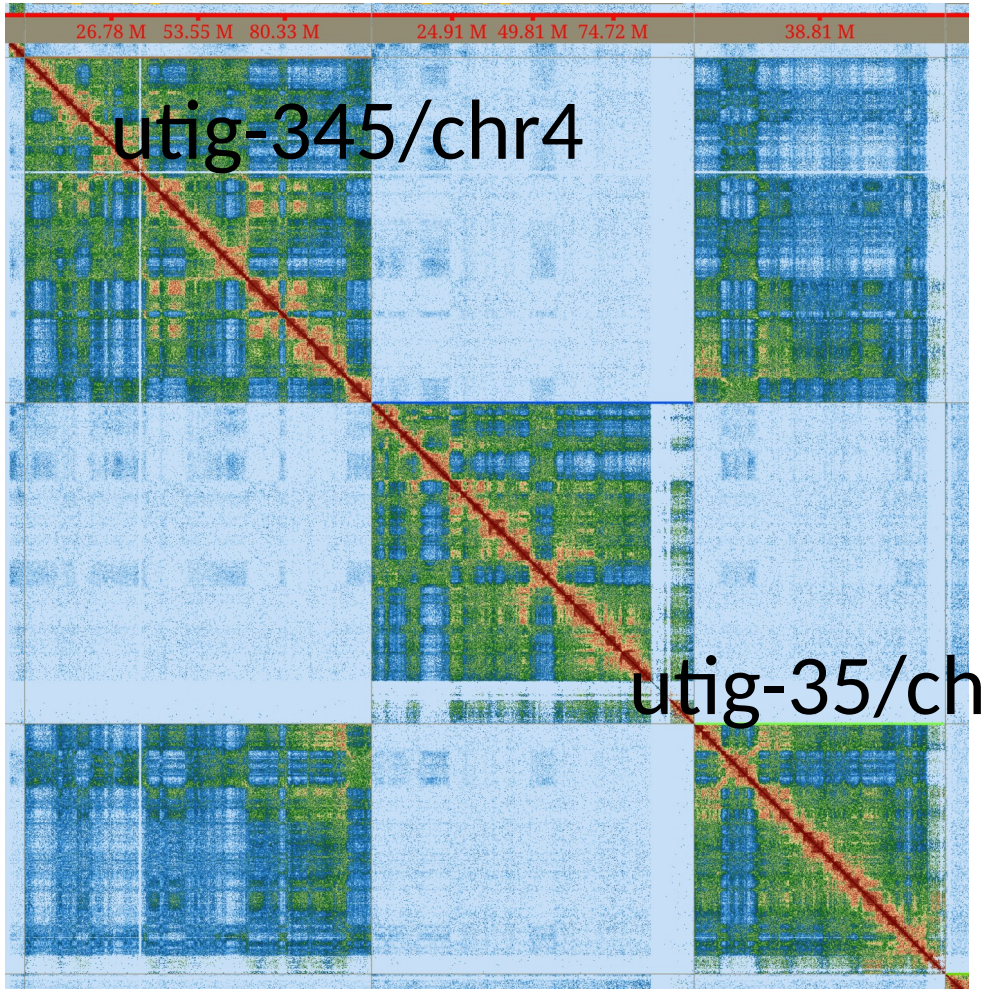
**Reference:
SHR-A3 in the NCBI

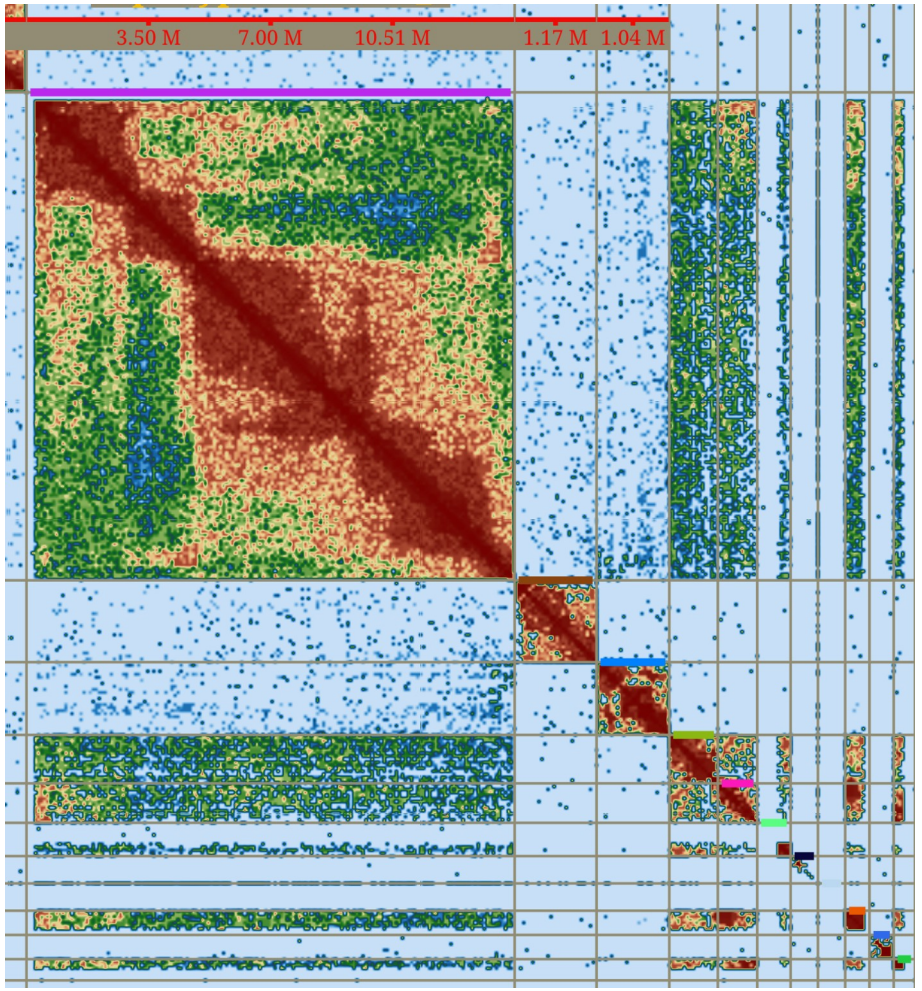
Visualizing assembly graphs

Bandage NG



Connecting Associated Scaffolds through Gaps





utig4-21/chr19

utig4-76

utig4-153

utig4-70

utig4-267

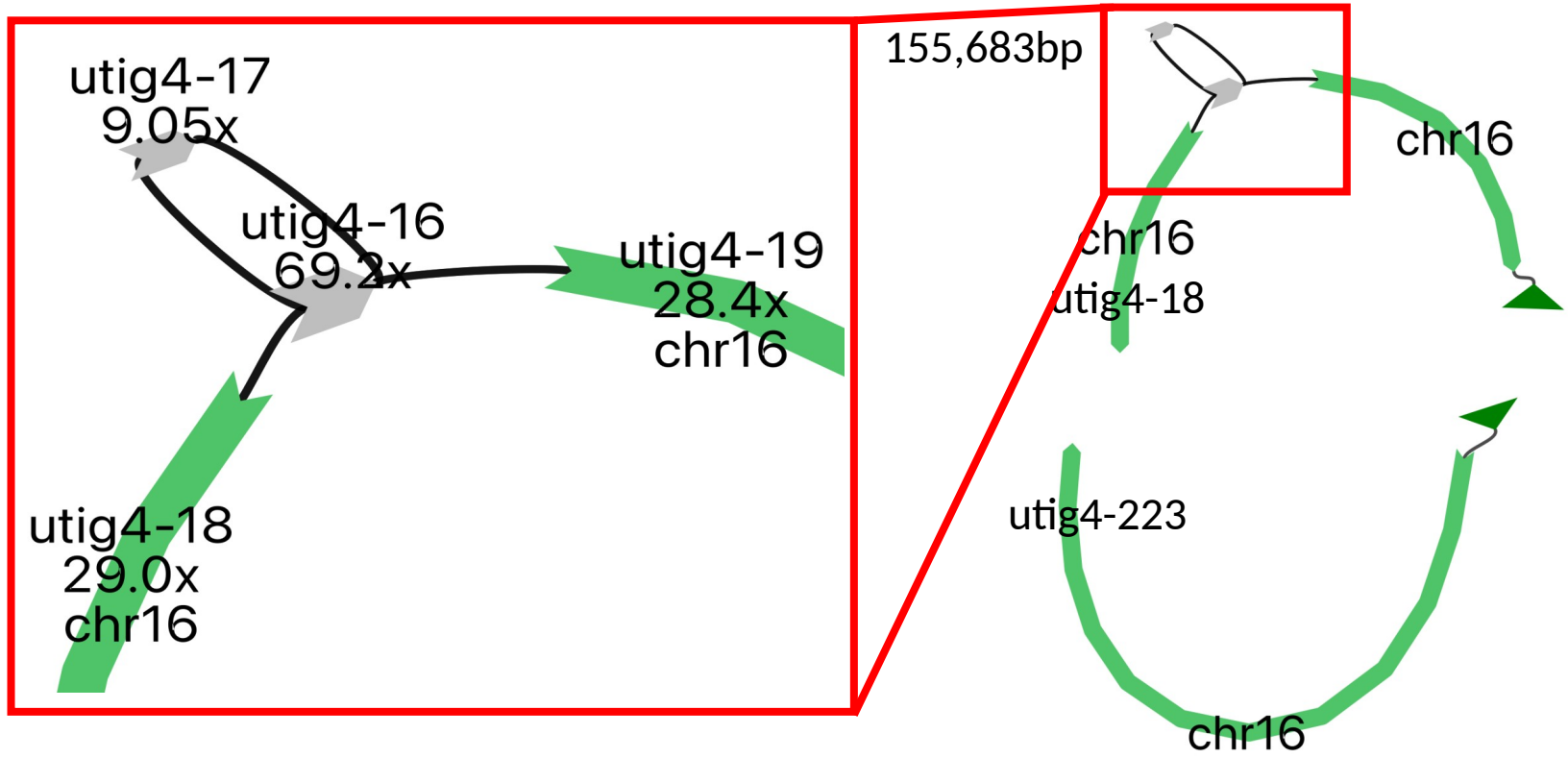
Querying aligned ONT Path

grep utig4-19 ont-aligned.gaf |grep utig4-18
NULL

grep utig4-18 ont-aligned.gaf |grep utig4-17
NULL

grep utig4-19 ont-aligned.gaf |grep utig4-17
<utig4-19<utig4-16>utig4-17
<utig4-17>utig4-16>utig4-19

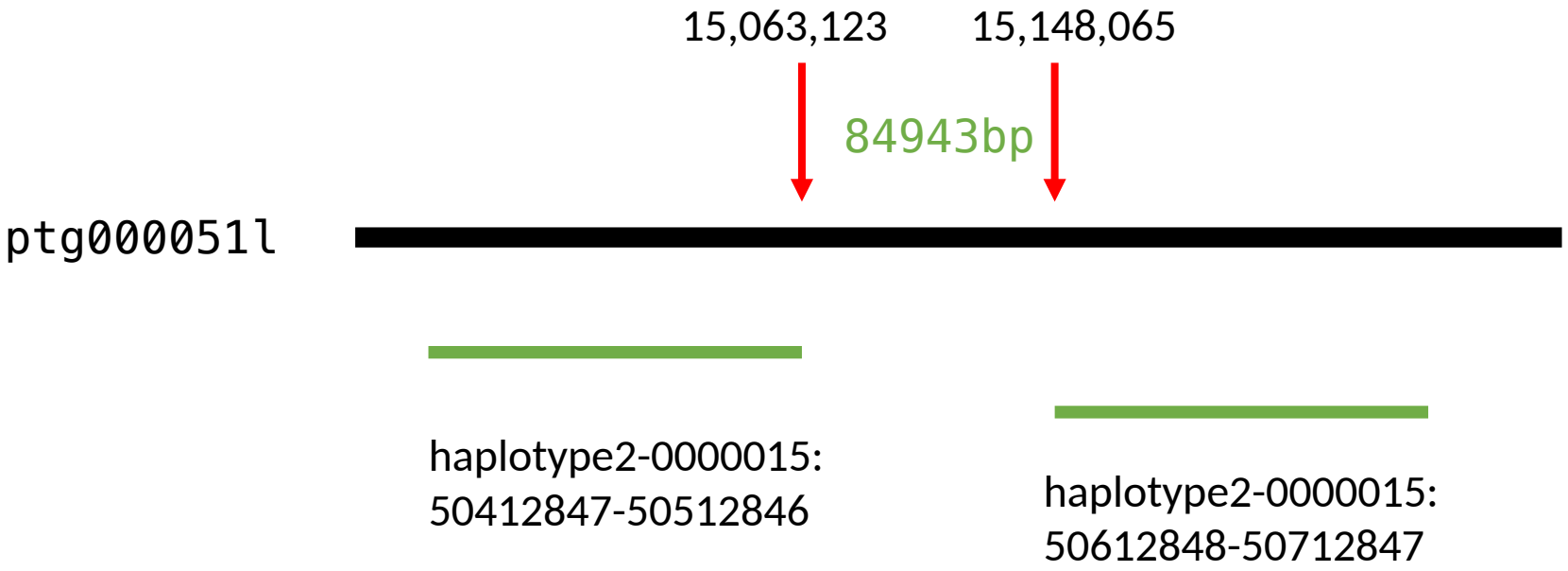
grep utig4-18 ont-aligned.gaf |grep utig4-16 |wc -l
66
grep utig4-19 ont-aligned.gaf |grep utig4-16 |wc -l
84
grep utig4-17 ont-aligned.gaf |grep utig4-16 |wc -l
157

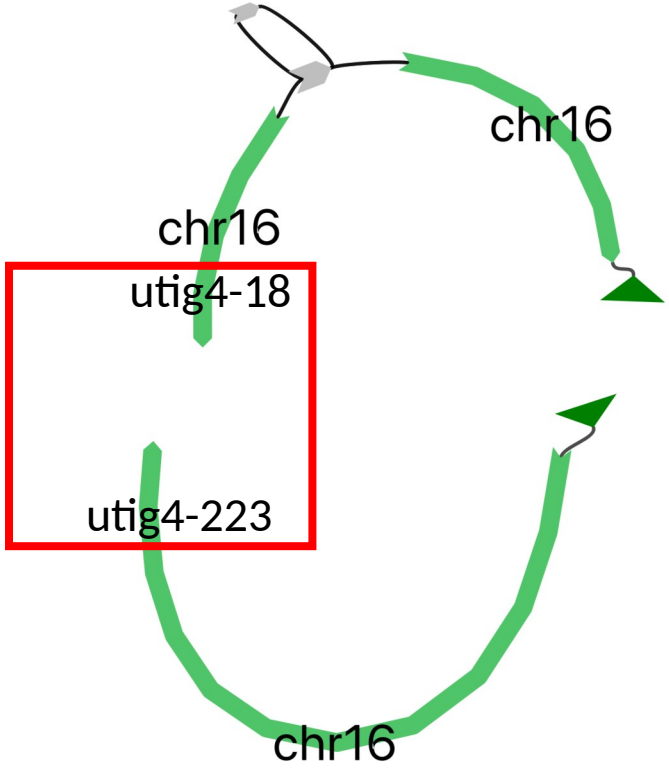


for each edge using once and 134kb utig4-16 coverage 69.2x
the real path `utig4-19-, utig4-16-, utig4-17+, utig4-16-, utig4-18+`

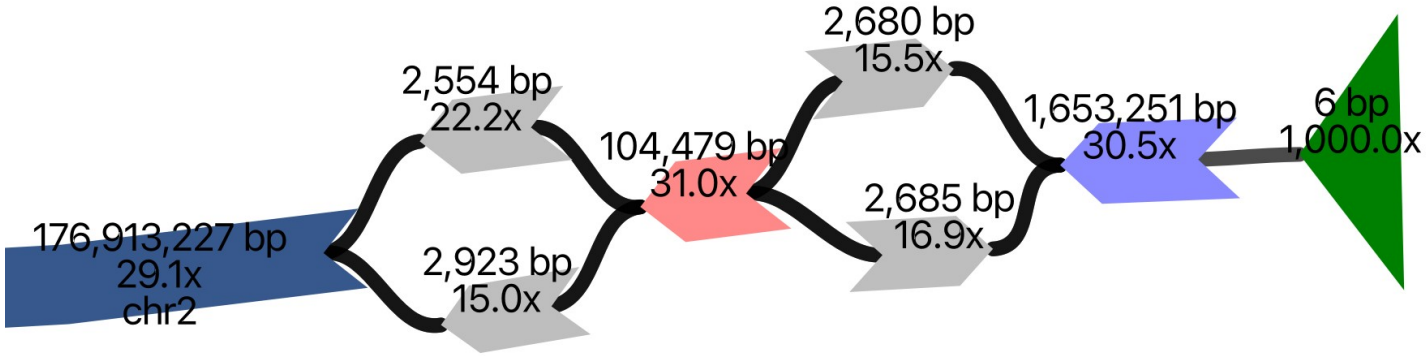
Gap Filling Using Assembly Results from hifiasm

The input data is exactly the same as the Verkko.
Different assemblers can complement each other's shortcomings through comparison and integration of their results.





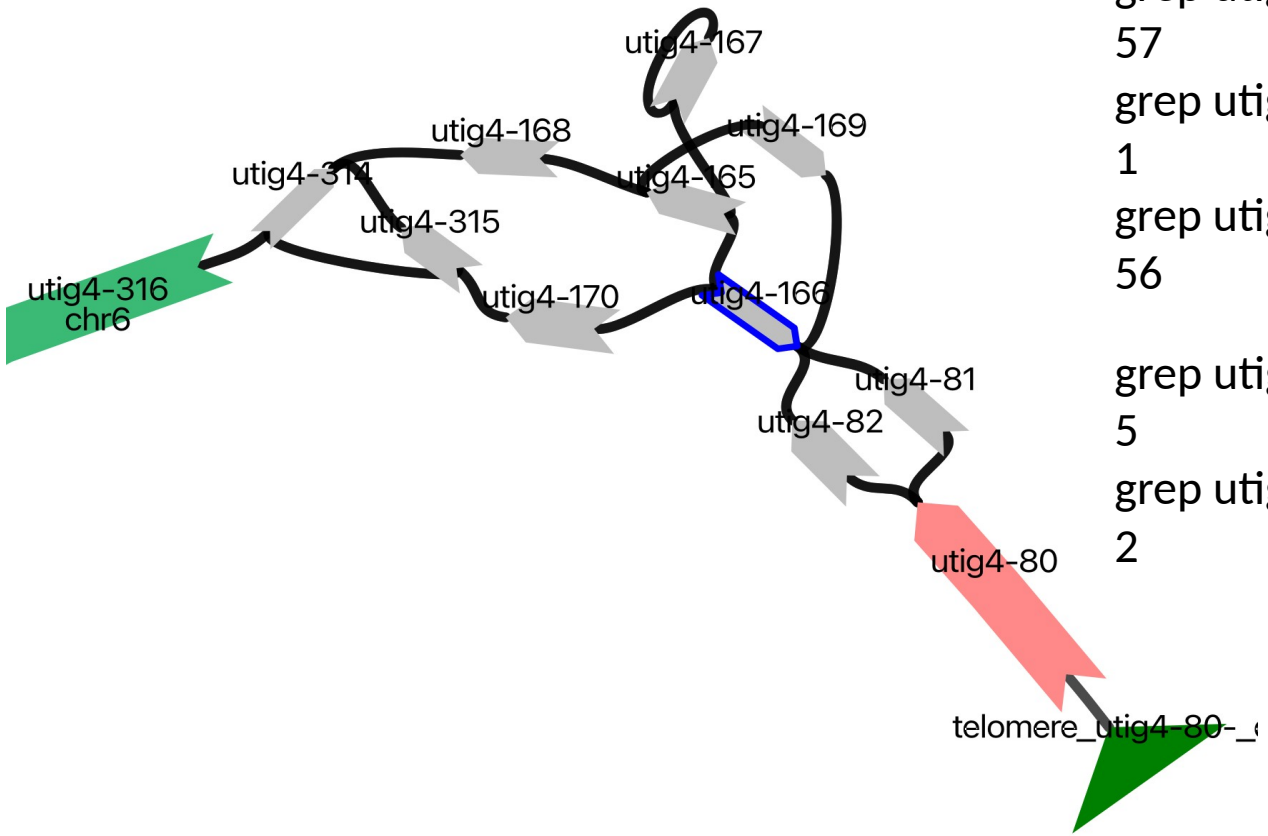
Simple long sequence repeats and heterozygous regions



grep utig4-202 ont-aligned.gaf |grep utig4-120

>utig4-202>utig4-342>utig4-342<utig4-120

Resolve complex tangles through critical paths

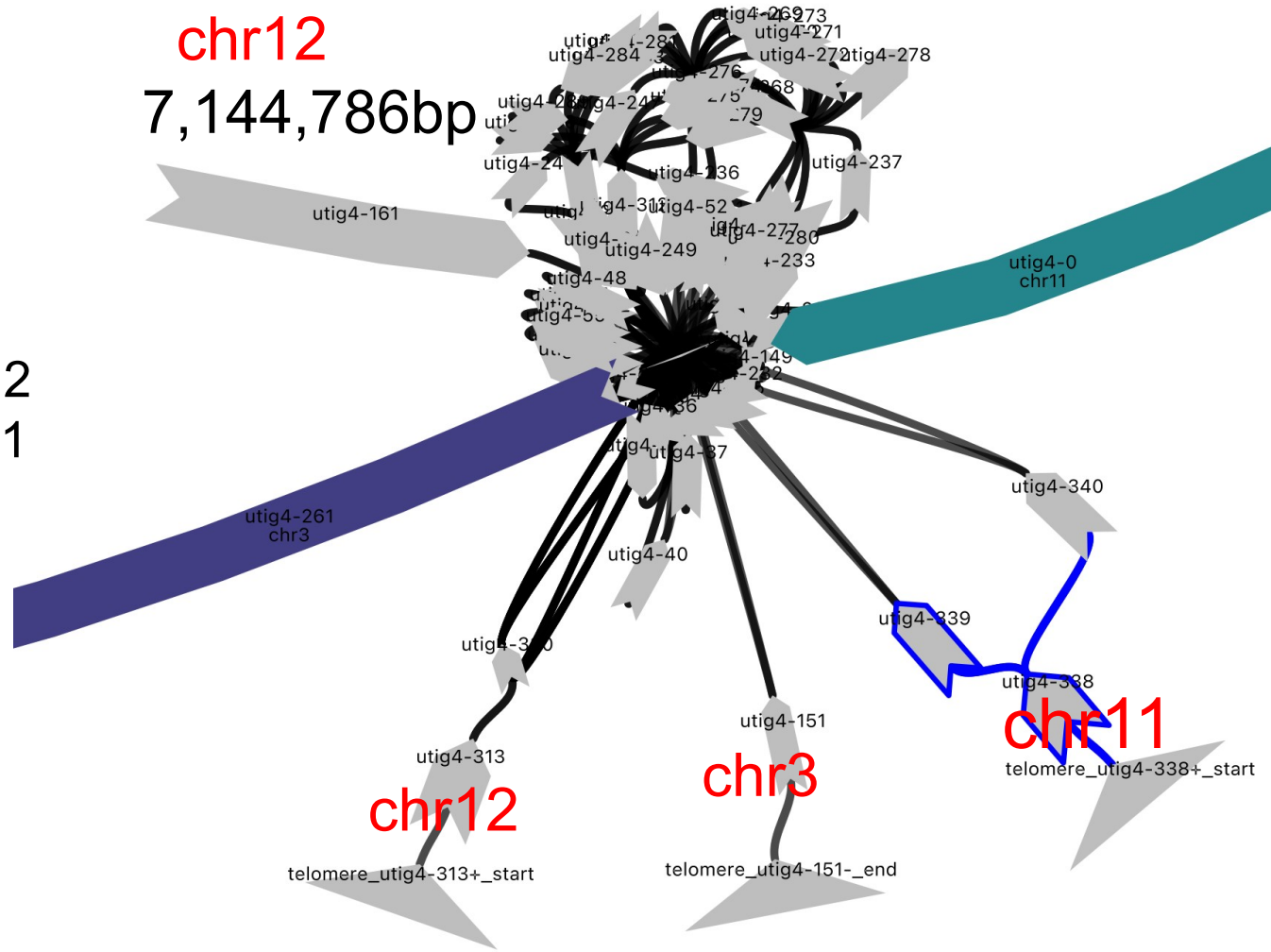



```
grep utig4-80 ont-aligned.gaf |grep utig4-166 |wc -l
57
grep utig4-80 ont-aligned.gaf |grep utig4-166 |grep utig4-81 |wc -l
1
grep utig4-80 ont-aligned.gaf |grep utig4-166 |grep utig4-82 |wc -l
56

grep utig4-166 ont-aligned.gaf |grep utig4-314 |grep utig4-315 |wc -l
5
grep utig4-166 ont-aligned.gaf |grep utig4-314 |grep utig4-168 |wc -l
2
```

rDNA arrays

utig4-151 -> haplotype2-0000027 -> chr3
utig4-313 -> unassigned-0000229 -> chr12
utig4-338 -> unassigned-0000244 -> chr11



Polishing:

To avoid over-polishing the NUMT region, the MT genome must be assembled before polishing.

MitoHiFi:

- map PacBio HiFi reads to the close-related mitogenome
- filter out any mapped reads that are larger than the reference mitogenome to avoid NUMTS
- hifiasm to assemble the mapped and filtered reads
- blast of the contigs versus the close-related mitogenome
- filtering BLAST output to select target sequences
- circularize, annotate and rotate each filtered contig

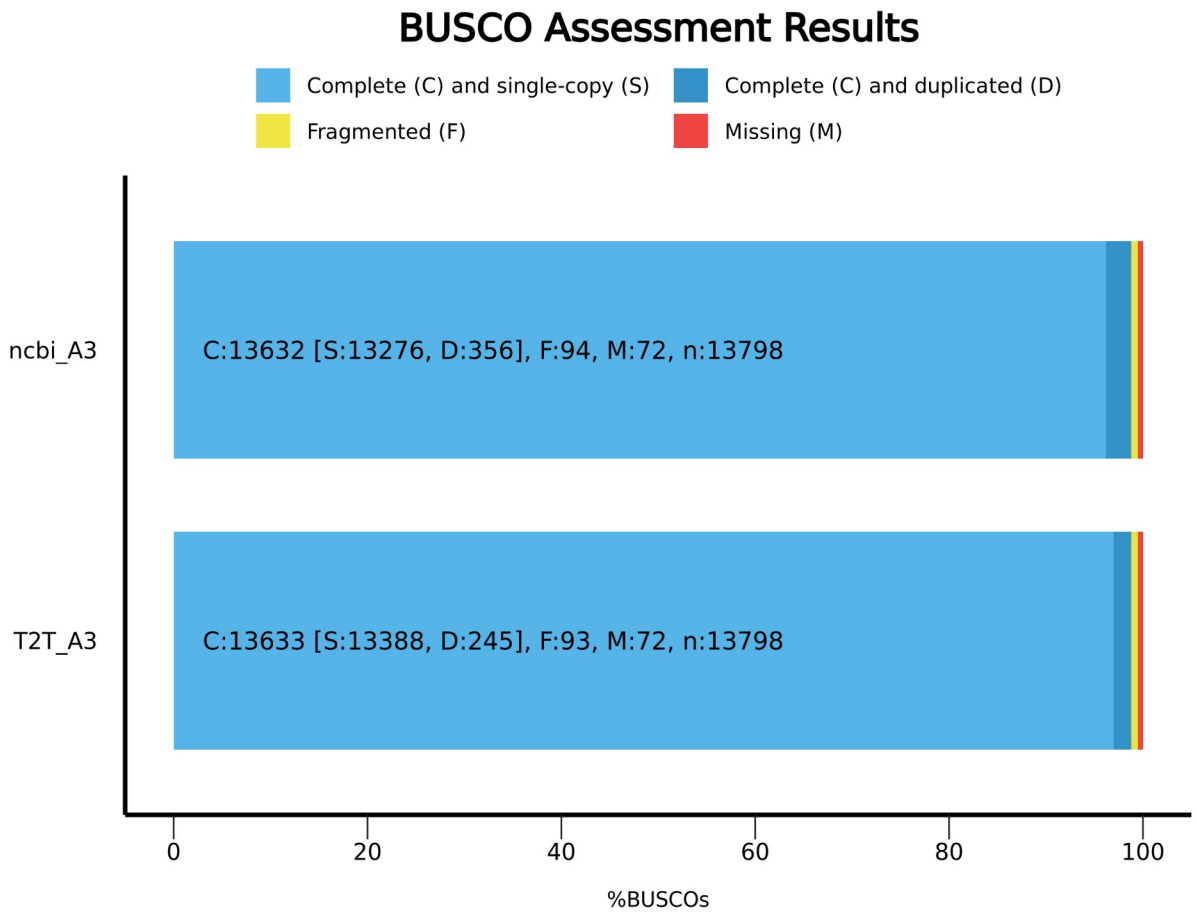
Deeppolisher && PacBio HiFi
Pilon && Illumina WGS

Assembly Comparison

	NCBI_SHRSP_A3	T2T_A3
N50	138,881,597	144,442,684
L50	8	8
Total_Length	2,907,517,304	2,852,523,581
Longest_contig	276,968,795	273,948,194
Number of Contigs	4,130	277
Number of Gaps	1,610	7
Length of Gaps(Kb)	1605.31	700.01
Number of telomeres	16	42
T2T_chromosomes	1	20 **

** chr7 & chrY has one end with telomere

Comparison evaluation



**lineage: glires_odb10

Compleasm:

	ncbi_A3	T2T_A3
S:	97.90%, 9032	98.52%, 9089
D:	1.97%, 182	1.35%, 125
F:	0.08%, 7	0.08%, 7
I:	0.00%, 0	0.00%, 0
M:	0.05%, 5	0.05%, 5
N:	9226	9226

**lineage: mammalia_odb10

Comparison evaluation

Merqury:

	Error Rate	Quality Value	Completeness
NCBI_A3	4.10962E-05	43.862	99.8118
T2T_A3	2.86592E-05	45.427	99.8148

Acknowledgements

- NHGRI
  - Sergey Koren, PhD
  - Adam M. Phillippy, PhD
- University of Louisville
  - Elizabeth Hudson, MS
  - Melissa Laird-Smith, PhD
- University of Kentucky
  - Theodore Kalbfleisch , PhD
- University of Texas McGovern School of Medicine
  - Peter Doris , PhD

NIH R01HG011252

Martin-Gatton
College of Agriculture,
Food and Environment

Thank you for your attention!

Questions?

kai.li@uky.edu