Genome assembly quality control

Bioplatforms Fungi Genomics Workshop 2024

Australian National University

Rita Tam

ritahltam

ritatam

“I have my first genome assembly (`.fasta`)! Now what?”

Evaluate assembly quality!

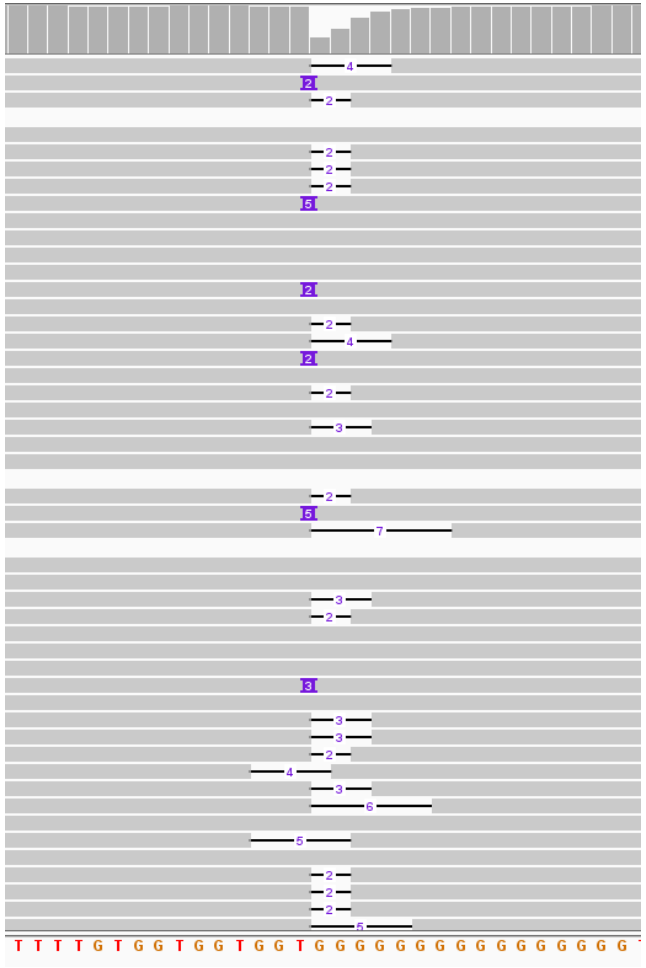
Never assume your genome assembly is “error-free”.

This is a “draft” assembly.

How assembly can fail perfection

Sequencing data

- Insufficient read depth
- DNA sequencing errors (biases)
  - Lower read quality towards read 3’ end
  - Prefers low-GC
  - Homopolymers (e.g. AAAAAAA or CCCCCC)
- Contaminant DNA
  - Bacteria, unexpected fungal species, host plant, etc.

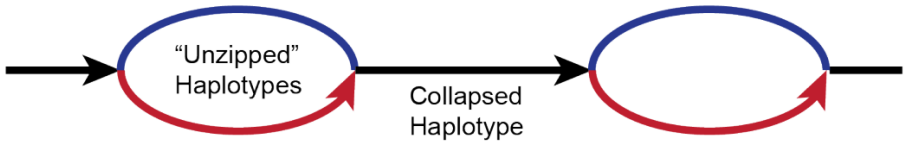


How assembly can fail perfection

Genome complexity

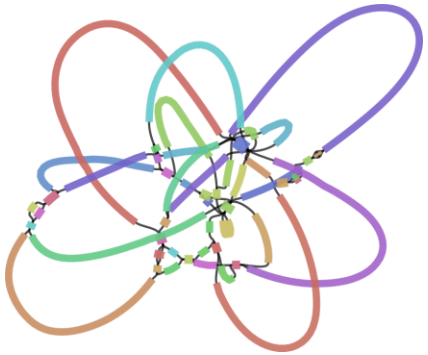
- Reads not long enough to span repeats
- Heterozygosity (if diploid)

    e.g. confusion at highly homozygous regions



Assembly algorithms

- Tolerance for sequencing errors
- Ploidy-awareness



https://www.pacb.com/wp-content/uploads/Kingan-ASHG-2018-FALCON-Phase-integrates-PacBio-and-HiC-data-for-de-novo-assembly.pdf

Why is it important to evaluate assembly quality?

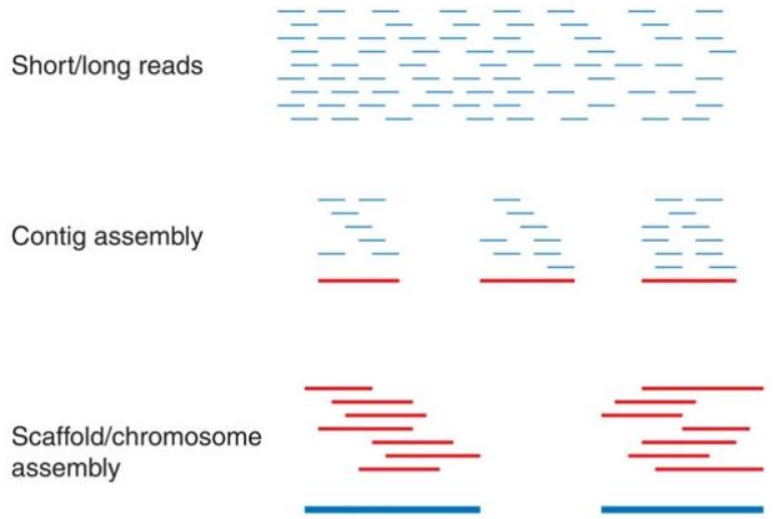
Know how trustworthy your genome assembly is!

If errors go unchecked, they can be

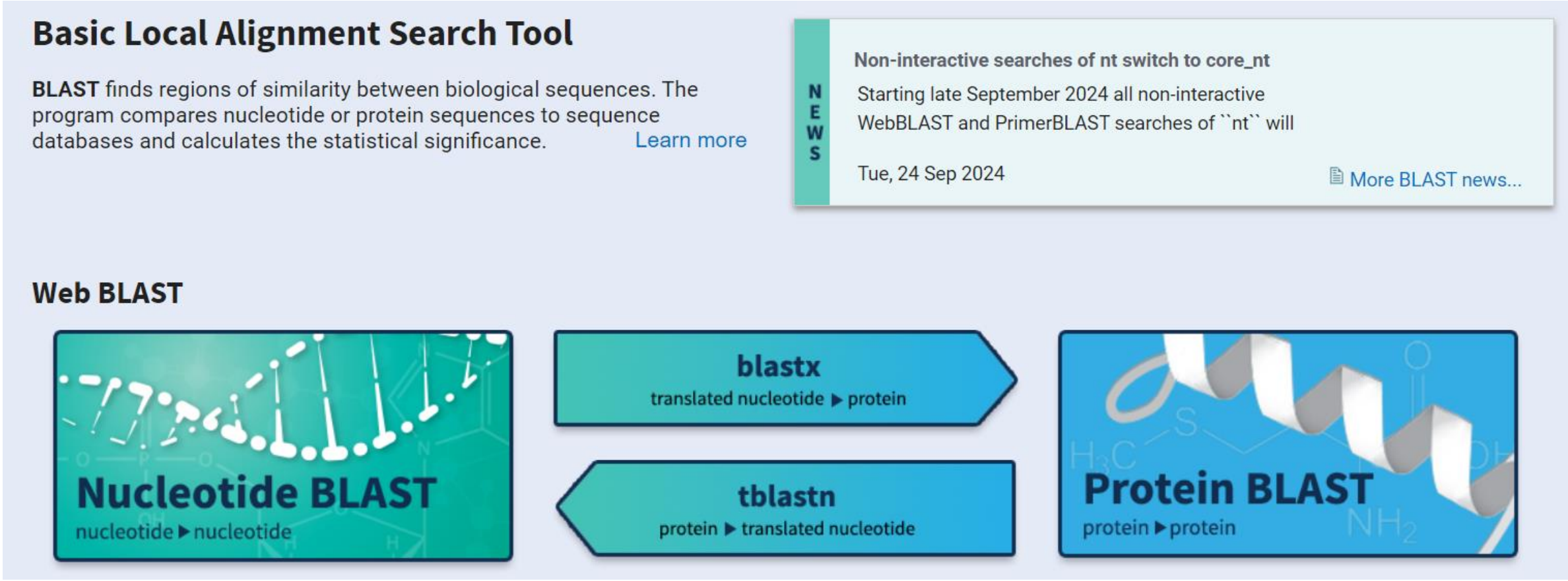
- propagated downstream
- misinterpreted as true biological events (e.g. a deletion)!

Users can make careful decisions

- revisit library prep?
- next steps?
  - scaffolding
  - curation
  - annotation

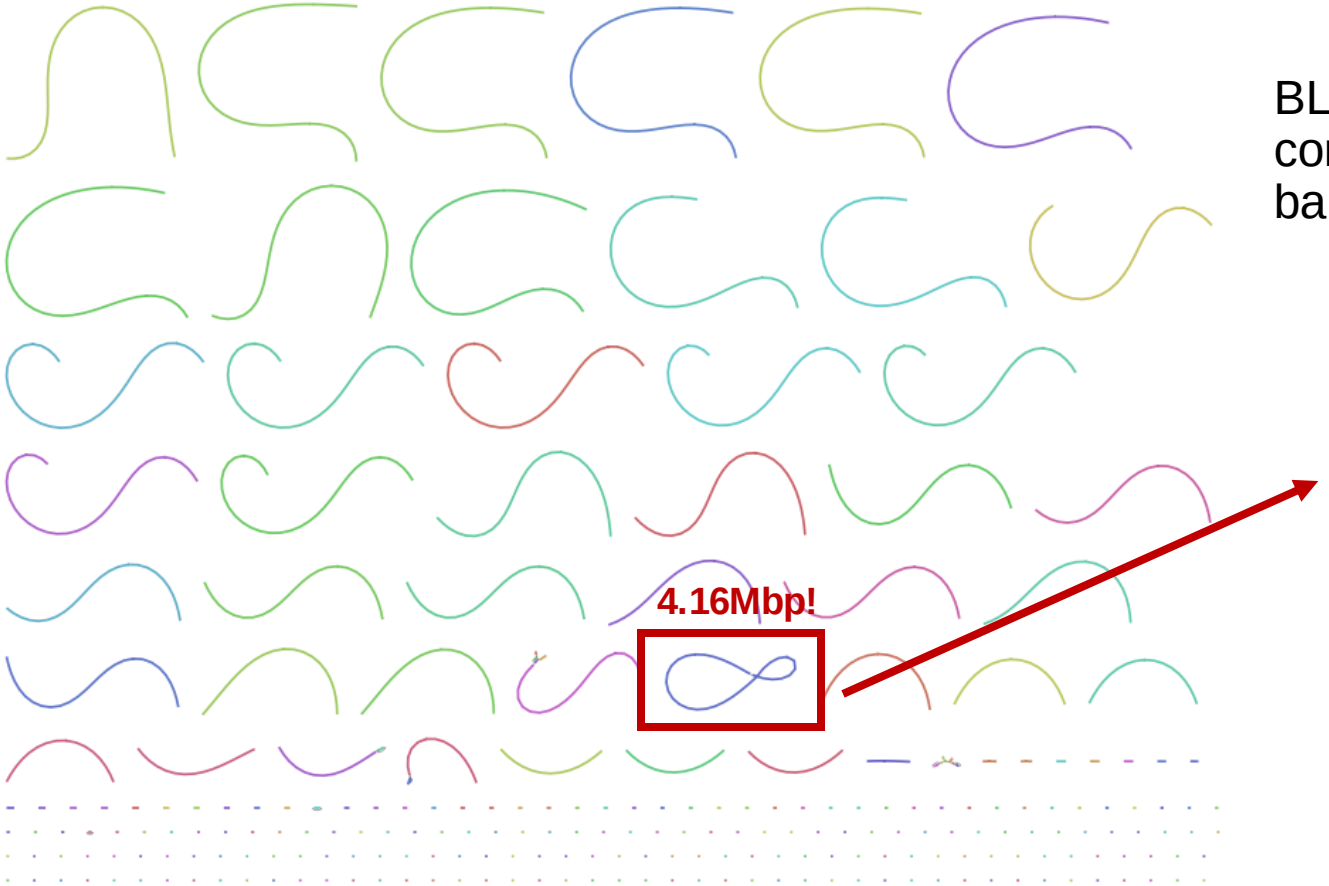


First – BLAST check your assembly!



First – BLAST check your assembly!

Puccinia striiformis f. sp. *tritici*
(wheat stripe rust fungi)



BLAST to detect and remove contaminant (e.g. wheat, bacteria) & mtDNA contigs.

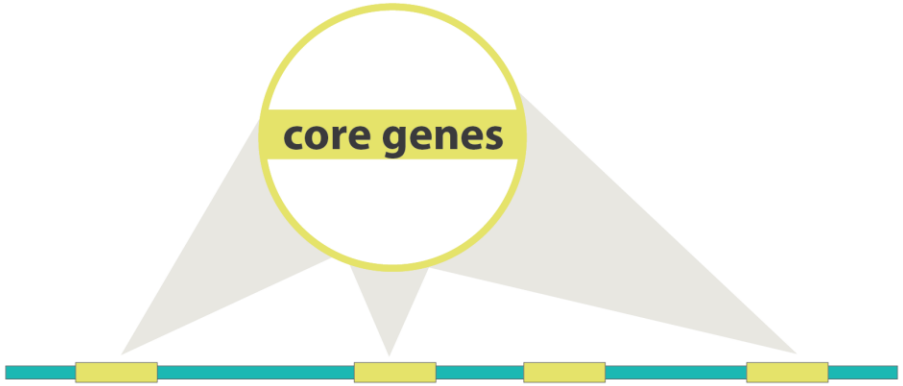
Herbaspirillum genome fully assembled along with the fungal genome!

Zahra Chew

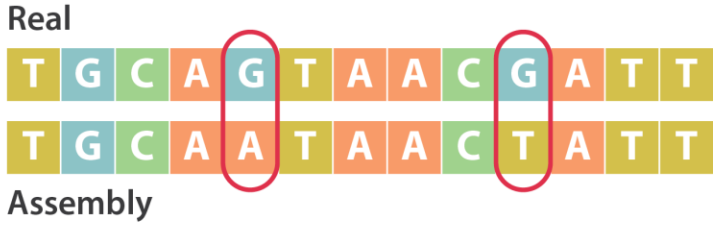
The “3C” rules for assmembly QC

Contiguity

Completeness



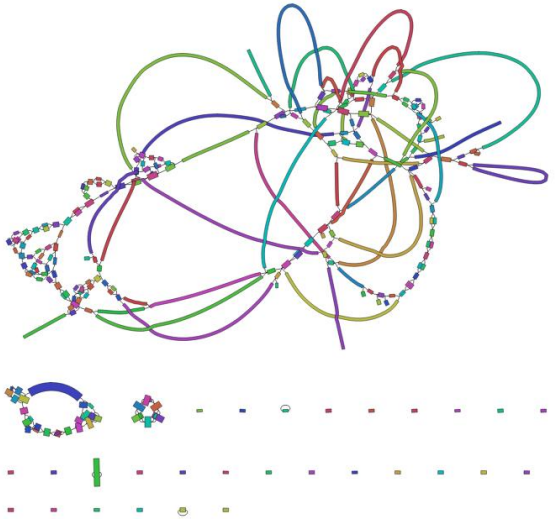
Correctness



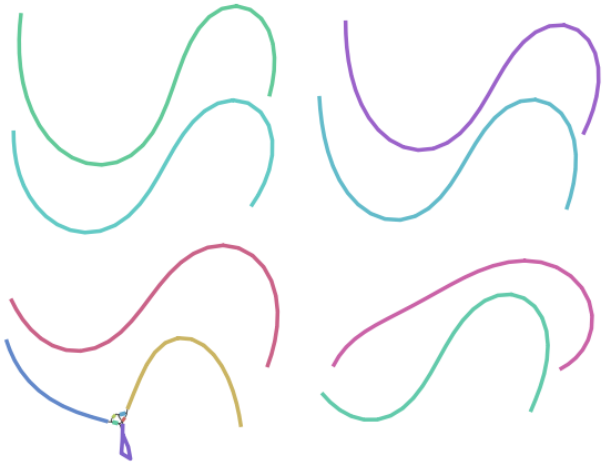
Contiguity

What we want:
Fewer and **longer** contigs,
chromosome-scale

Fragmented



Chromosome-scale



N50/L50

N50

Sequence length of the shortest contig at 50% of the assembly length.

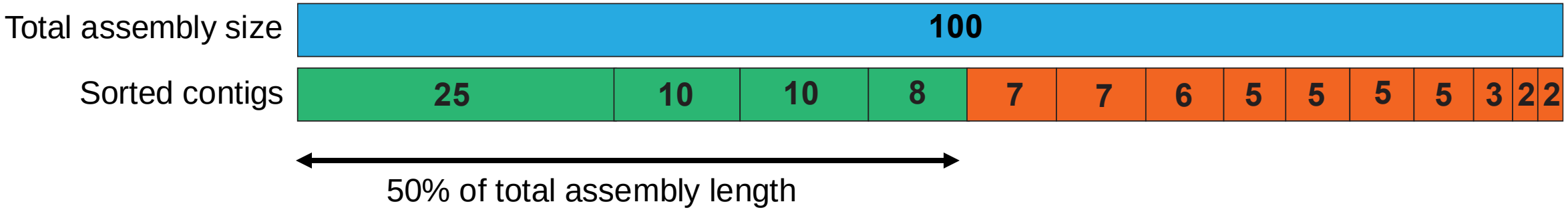
Generally, the higher the better.

L50

Smallest count of contigs whose length sum make up 50% of the assembly.

Generally, the lower the better.

N50/L50



Total assembly size = 100

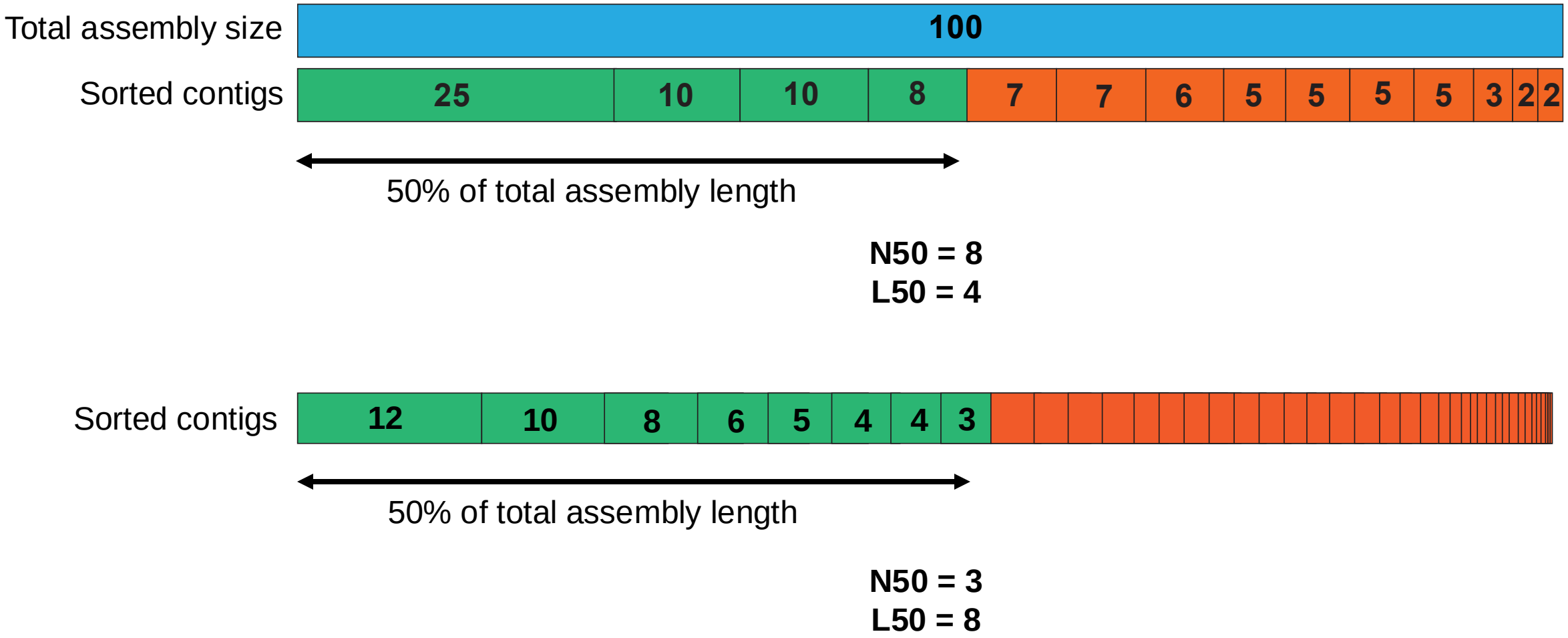
Sort contigs by length

50% of total length is contained within sequences of at least 8 (25+10+10+8 = 53, >= 50)

N50 = 8

L50 = 4

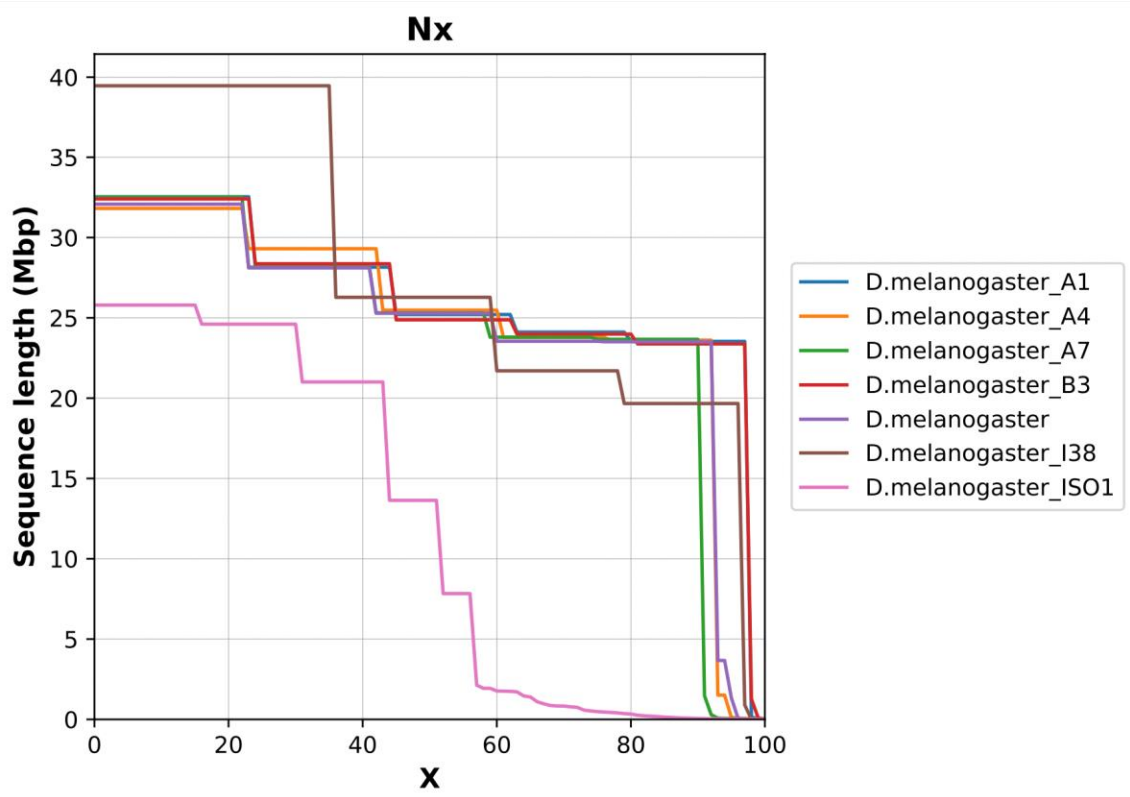
N50/L50

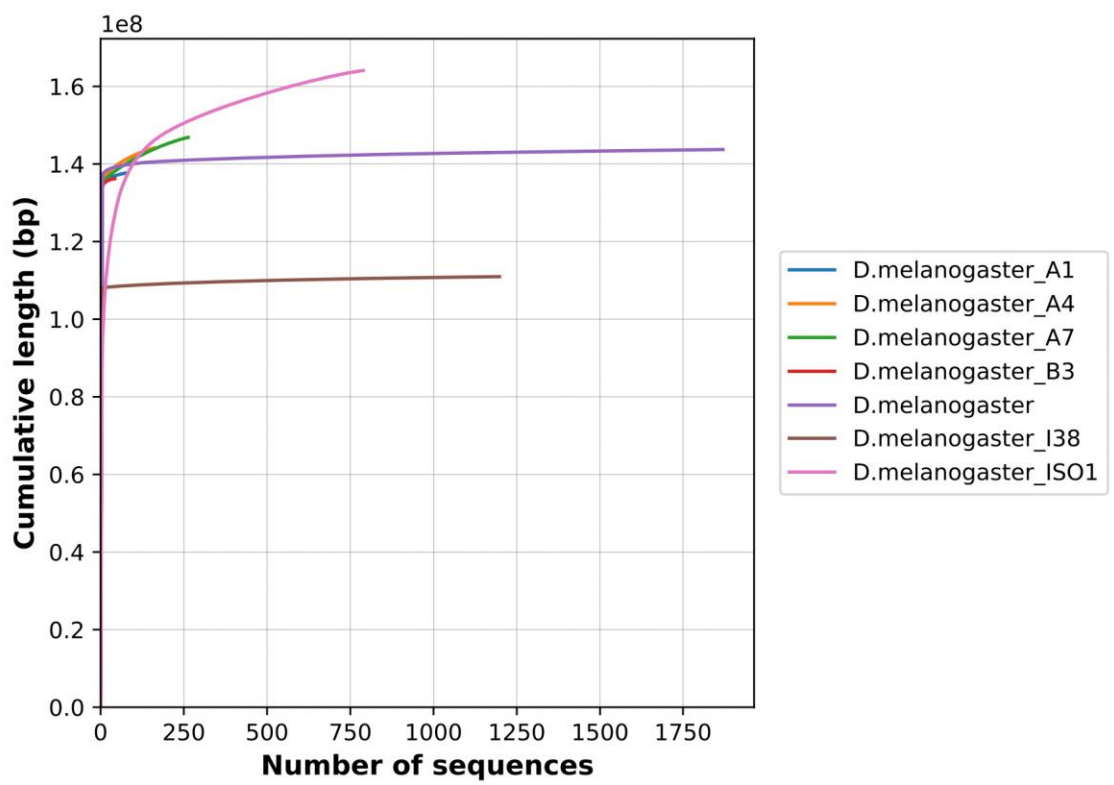

Total assembly size
100
Sorted contigs
25
10
10
8
7
7
6
5
5
5
5
3
2
2
50% of total assembly length
N50 = 8
L50 = 4
Sorted contigs
12
10
8
6
5
4
4
3
50% of total assembly length
N50 = 3
L50 = 8

Nx curve

QUAST

Quality Assessment Tool for Genome Assemblies by CAB





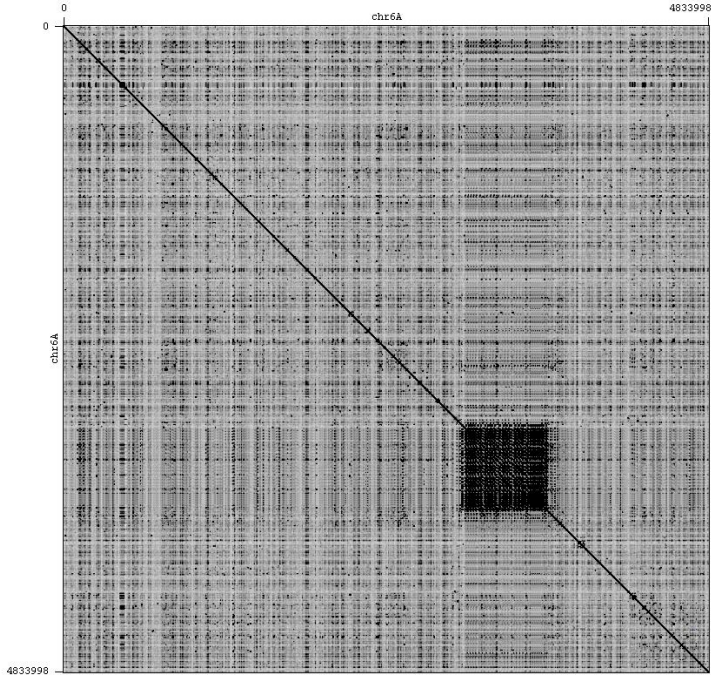
Contiguity measured at repeats

- LTR retrotransposons are abundant in fungi (Muszewska et al. 2017)
- Assembly contiguity can be measured at repeats, e.g. LTR transposons

LTR Assembly Index (LAI)
(Ou et al. 2018)

Category	LAI
Draft	0 ≤ LAI < 10
Reference	10 ≤ LAI < 20
Gold	20 ≤ LAI

LTR retrotransposon-dominated region near *P. striiformis* f.sp. *tritici* mating type locus



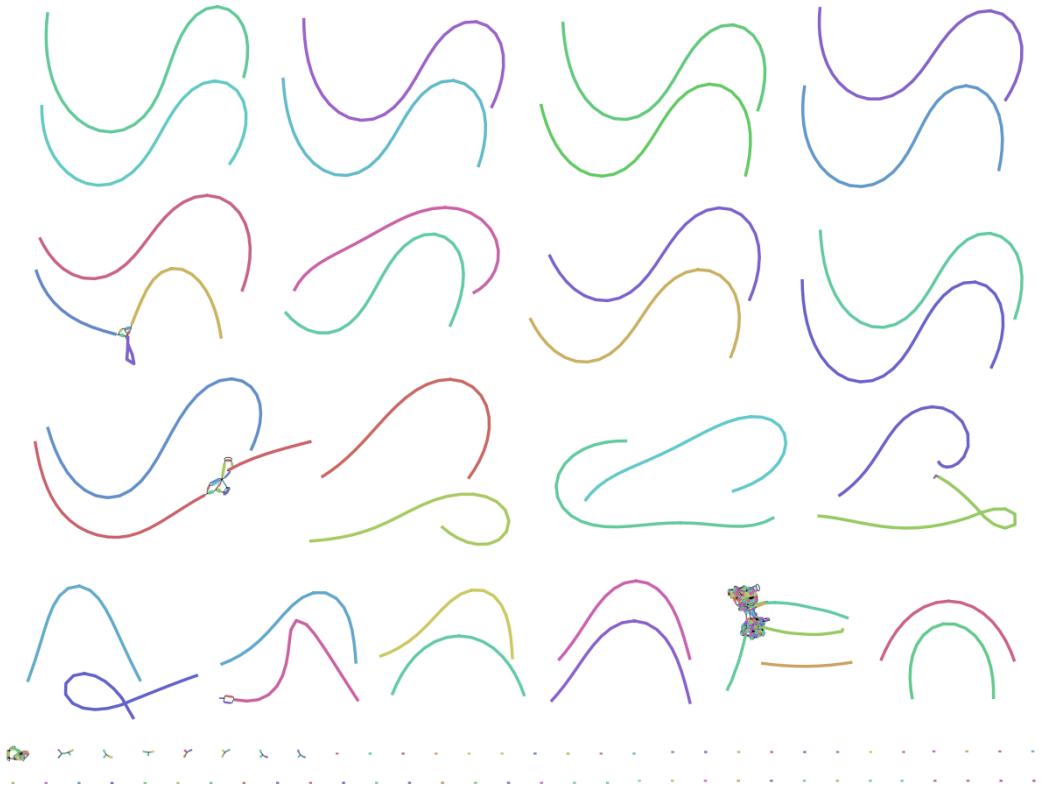
Limitations

- Less suitable for species with low LTR abundance
- Intrinsic degradation rate of LTRs varies between species
- **General guide only, low LAI score =/= bad!**

Genome graph

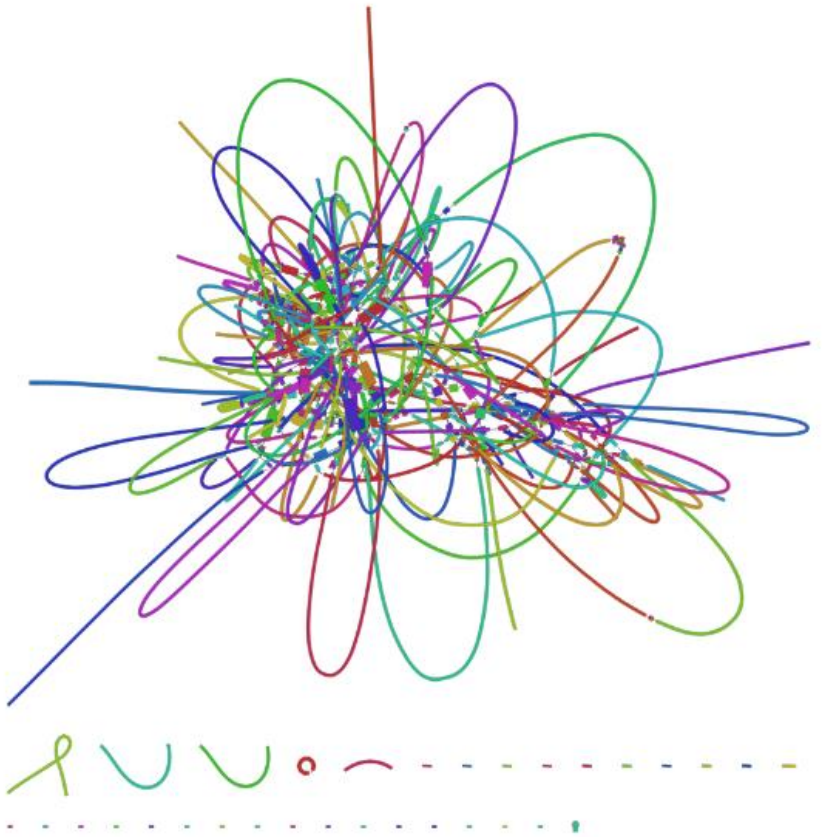
Quick visualisation for a “feeling” of contiguity

assembly_output.fasta
assembly_output.gfa → Bandage
…



Good sign:

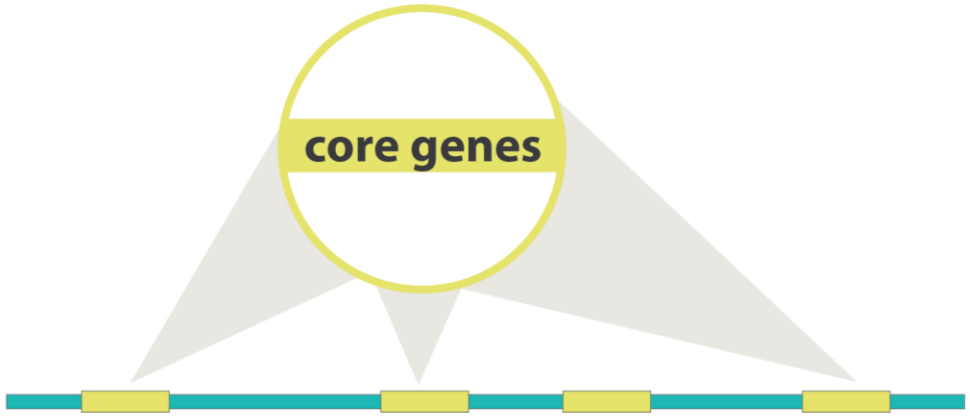
- Chromosome-scale contigs



Not so good sign:

- Chromosomes broken into small contigs
- Ambiguous contig connections

Completeness



Assembly size

Gene space completeness

- BUSCO
- Reference-based

Telomere counting

Assembly gap

Assembly size vs Expected

$$\frac{Assembly\ size}{Expected\ genome\ size}$$

Compare the assembly size to expected genome size

Very rough estimation

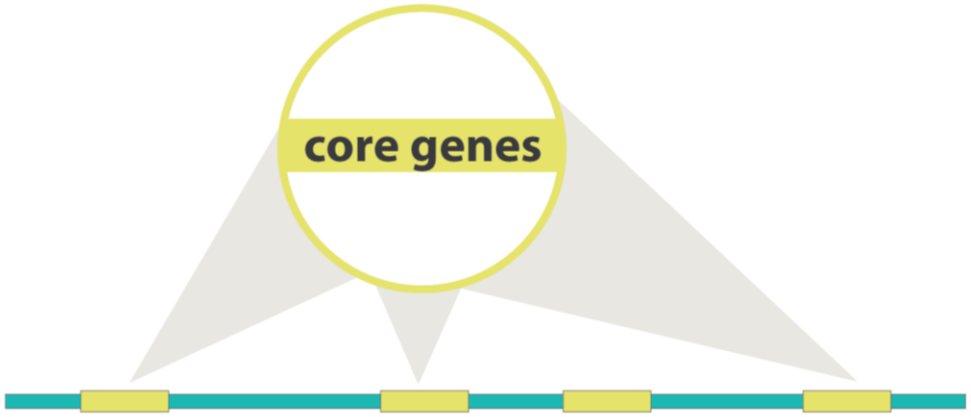
Limitations

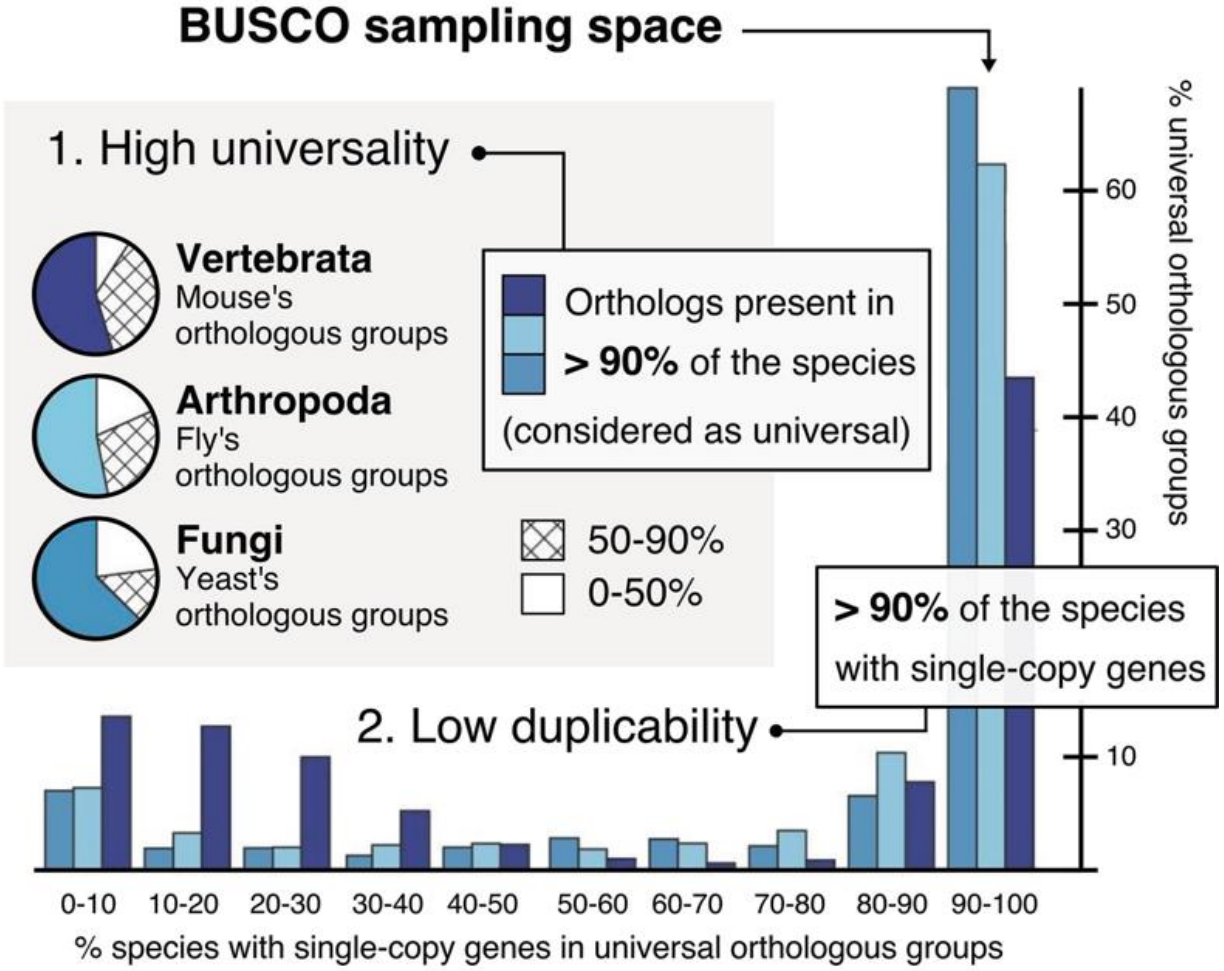
- Empirical evidence can have inaccuracies
- Genome size can vary between species, even genotypes
- Contaminant DNA can inflate it

Gene space completeness - **BUSCO**

BUSCO: Benchmarking Universal Single-Copy Orthologs

Evaluate “core” gene content using a predefined set of highly conserved orthologs





Gene space completeness - **BUSCO**

```
lineages
├── agaricales_odb10
├── agaricomycetes_odb10
├── archaea_odb10
├── ascomycota_odb10
├── bacillales_odb10
├── bacteria_odb10
├── endopterygota_odb10
├── eudicots_odb10
├── eukaryota_odb10
├── fungi_odb10
├── hemiptera_odb10
├── insecta_odb10
├── mammalia_odb10
├── metazoa_odb10
├── methanococcales_odb10
├── natrialbales_odb10
├── primates_odb10
├── saccharomycetes_odb10
└── solanales_odb10
```

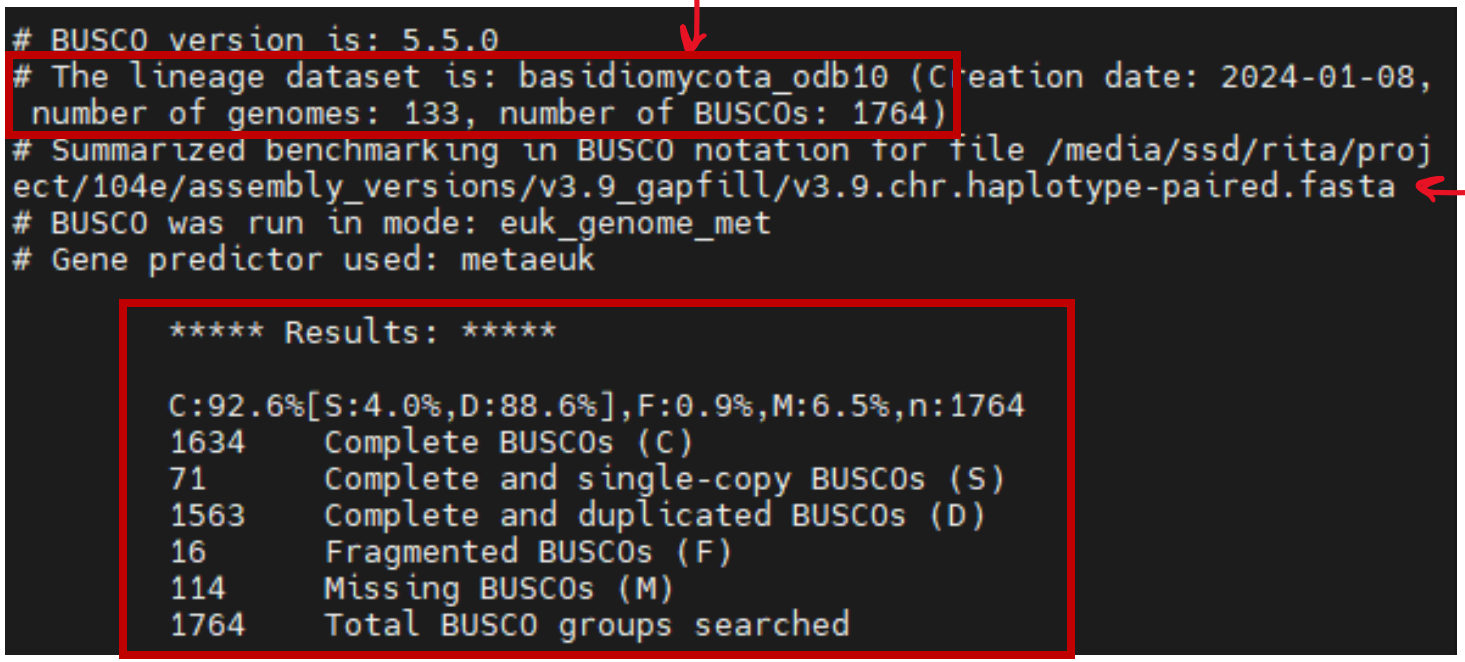
Fungi datasets Odb10

Kingdom: Fungi Odb10

Phylum level:	Class level:	Order level:
Ascomycota	Agaricomycetes	Agaricales
Basidiomycota	Dothideomycetes	Boletales
Microsporidia	Eurotiomycetes	Capnodiales
Mucoromycota	Leotiomycetes	Chaetothyriales
	Saccharomycetes	Eurotiales
	Sordariomycetes	Glomerellales
	Tremellomycetes	Helotiales
		Hypocreales
		Mucorales
		Onygenales
		Pleosporales
		Polyporales

Gene space completeness - **BUSCO**

BUSCO lineage dataset
for Basidiomycota



input fungal genome
assembly (.fasta)

BUSCO completeness results

Gene space completeness - **BUSCO**

BUSCO limitations

- “Completeness” inferred from a small gene subset
- Some species can lose conserved genes and yield lower BUSCO score, even if well-assembled
- Genes with long introns are hard to detect
- Better option: annotate genes with RNA-seq evidence, then assess BUSCO

Gene space completeness – **reference-based**

Reference-based

If previously annotated genes are available,
can count full and partial genes in your draft assembly

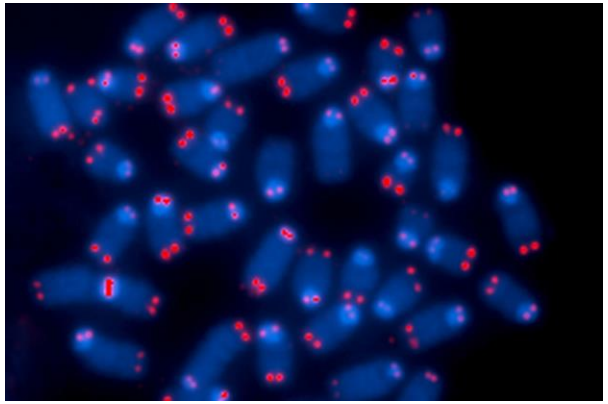
QUAST gene counting output

	assembly	genes	partial genes
draft assembly V1	verkko_duplex_assembly	31032	6
draft assembly V2	verkko_herro_assembly	31058	2

Yunshen Liu

Telomeres

- “Endpoints” of a chromosome assembly
- Telomere-to-telomere assembly now highly achievable, thanks to long reads



Article | Open access | Published: 30 May 2024

Complete telomere-to-telomere genomes uncover virulence evolution conferred by chromosome fusion in oomycete plant pathogens

Zhichao Zhang, Xiaoyi Zhang, Yuan Tian, Liyuan Wang, Jingting Cao, Hui Feng, Kainan Li, Yan Wang, Suomeng Dong, Wenwu Ye & Yuanchao Wang

RESOURCE ANNOUNCEMENT

OPEN

A Telomere-to-Telomere Genome Assembly Resource of *Bipolaris sorokiniana*, the Fungal Pathogen Causing Spot Blotch and Common Root Rot Diseases in Barley and Wheat

Yueqiang Leng, Yang Du, Jason Fiedler, Sajeet Haridas, Igor V. Grigoriev, and Shaobin Zhong

Affiliations

Published Online: 17 Jan 2024 | https://doi.org/10.1094/PHYTOFR-08-23-0108-A

Article | Open access | Published: 14 September 2022

Telomere-to-telomere genome sequence of the model mould pathogen *Aspergillus fumigatus*

Paul Bowyer, Andrew Currin, Daniela Delneri & Marcin G. Fraczek

Nature Communications **13**, Article number: 5394 (2022) | Cite this article

Telomere counting

JanaSperschneider / **FindTelomeres**

Assembled contigs

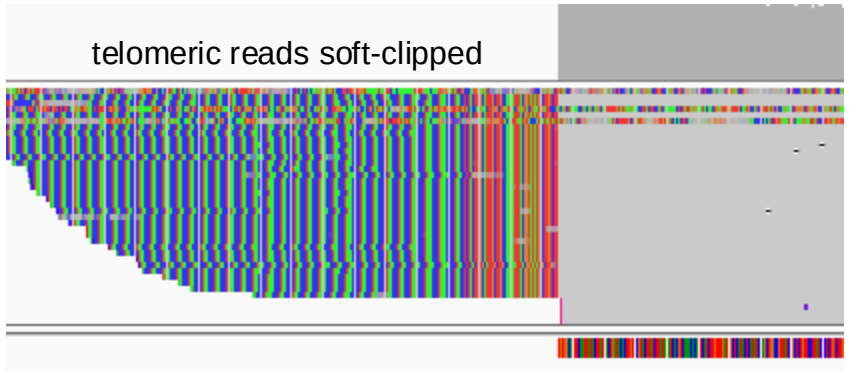
```
##########
356 sequences to analyze for telomeric repeats (TTAGGG/CCCTAA) in file assembly.raw.fasta
##########

haplotype1-0000001      Forward (start of sequence)     CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
haplotype1-0000001      Reverse (end of sequence)       GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTTAGGGTTAGG
haplotype1-0000002      Forward (start of sequence)     CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
haplotype1-0000002      Reverse (end of sequence)       GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000003      Forward (start of sequence)     ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000003      Reverse (end of sequence)       AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
haplotype1-0000004      Reverse (end of sequence)       GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000005      Forward (start of sequence)     CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
haplotype1-0000006      Forward (start of sequence)     ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000006      Reverse (end of sequence)       GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000007      Forward (start of sequence)     AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
haplotype1-0000007      Reverse (end of sequence)       TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAAGGGT
haplotype1-0000008      Forward (start of sequence)     CTAAACCCTAAACCCTAACCCTAAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000008      Reverse (end of sequence)       TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
haplotype1-0000009      Forward (start of sequence)     CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
haplotype1-0000009      Reverse (end of sequence)       GTTAGGGTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000010      Forward (start of sequence)     CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
haplotype1-0000010      Reverse (end of sequence)       GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000011      Forward (start of sequence)     ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000012      Forward (start of sequence)     ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACCCTAAACCCTA
haplotype1-0000014      Forward (start of sequence)     CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACC
haplotype1-0000014      Reverse (end of sequence)       GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
haplotype2-0000022      Forward (start of sequence)     CCTAACCCTAAACCCTAACCCTAACCCTAAACCCTAACCCTAACCCTAAC
haplotype2-0000022      Reverse (end of sequence)       GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
haplotype2-0000023      Forward (start of sequence)     CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
haplotype2-0000023      Reverse (end of sequence)       GGGTTAGGGTTAGGGTTAGGGTTAGGGTATTAGGGTTAGGGACTTGAATT
haplotype2-0000025      Forward (start of sequence)     CCCTAACCCTAACCCTAACCCTAACCCTCTAAACCCTAACCCTAACCCTA
haplotype2-0000025      Reverse (end of sequence)       GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
haplotype2-0000026      Forward (start of sequence)     TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA

Telomeres found: 64 (32 forward, 32 reverse)
```

Trap: a problematic read causing extension beyond telomere

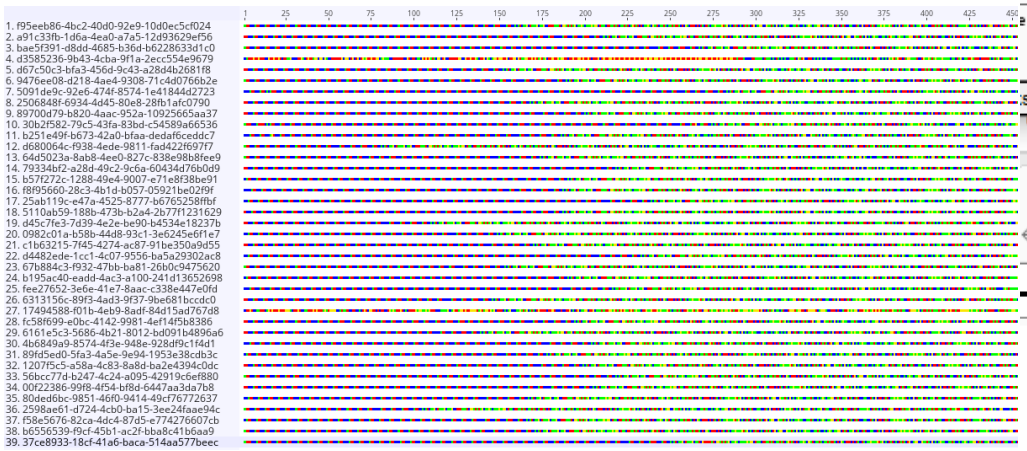
Actual telomere

Where telomere counter wants to find telomere, but failed

Recover telomeres “hidden” in the reads

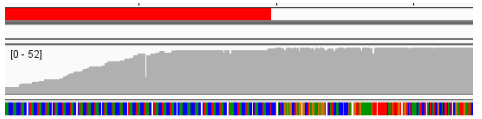


extract soft-clipped alignments with telomeric motif “TTAGGG/CCCTAA”

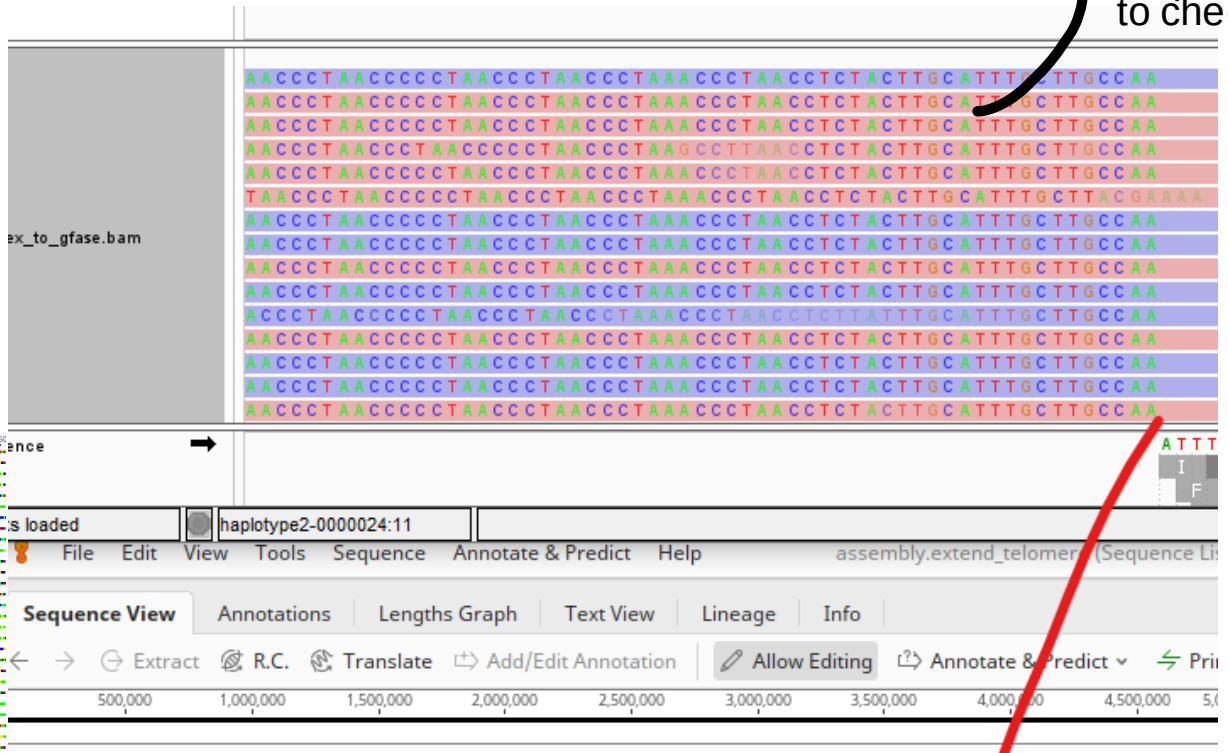


Local assembly (e.g. Flye)

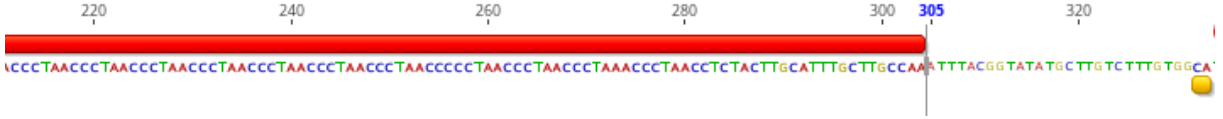
telomere consensus
(+ adjacent loci for anchoring)



map reads back again to check coverage



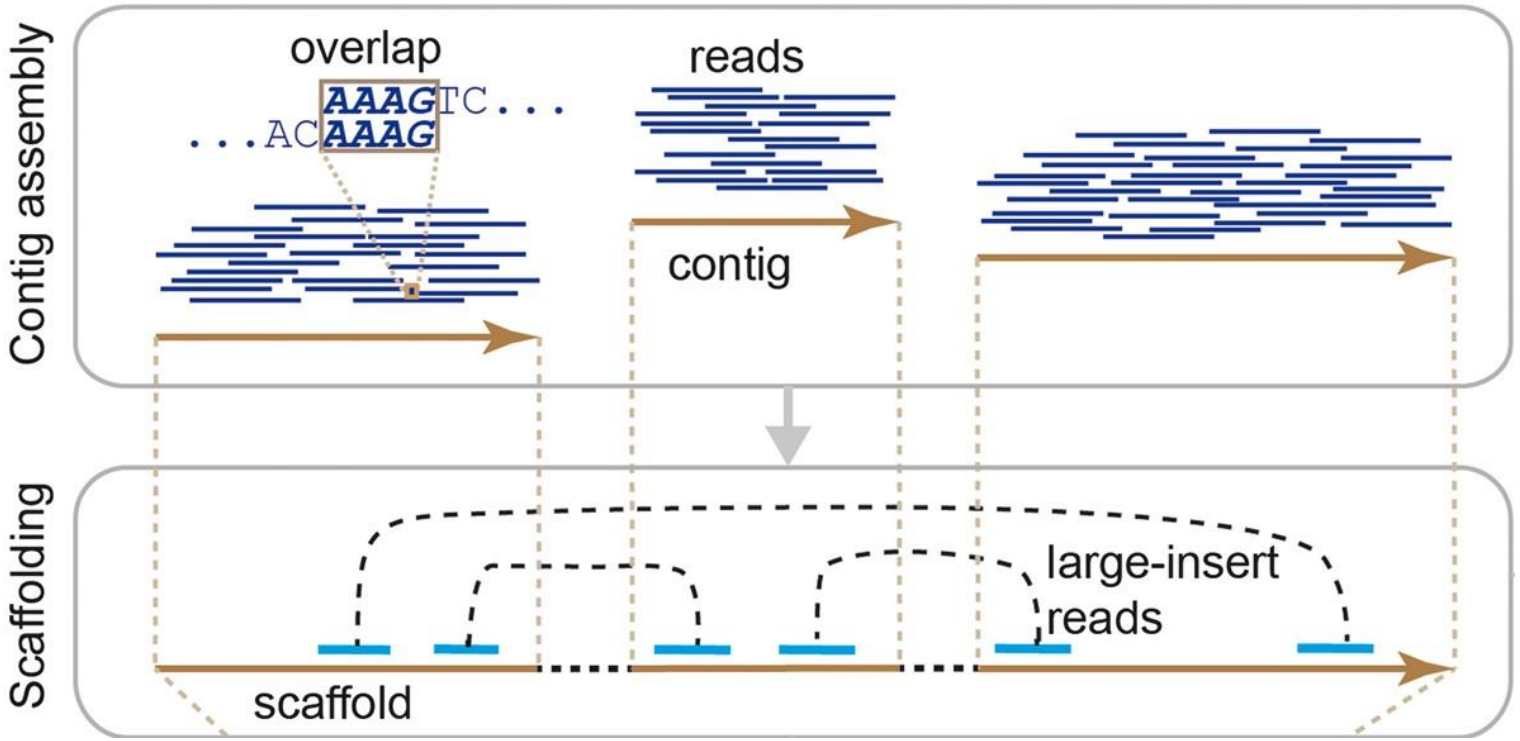
manually stitch telomeres back



Assembly gap

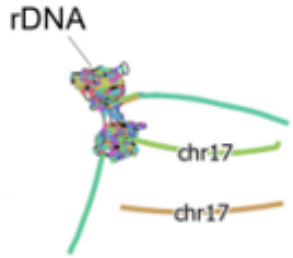
"unknown" nucleotides

- Assembly gap (Ns) should be introduced after scaffolding (not covered in this workshop)
- But there are assemblers that might do some for you (e.g. verkko)



Gaps between contigs → filled with NNNNNN…

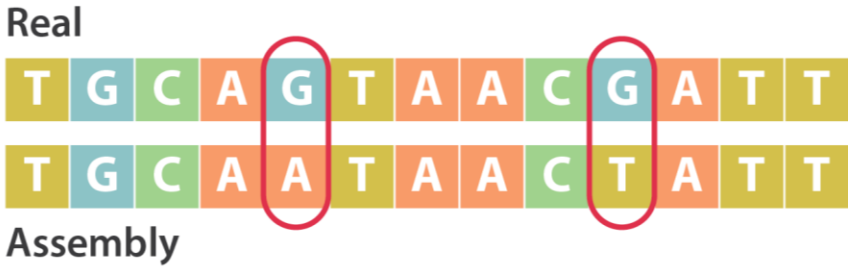
Assembly gap at ribosomal DNA array



Highly repetitive tandem repeats, identical copies

ATCCATCGCAATCGAATAGGNNNNNNNNNNNNNNNNCATTTACCGACGATACACAG

Correctness



SNP/indels

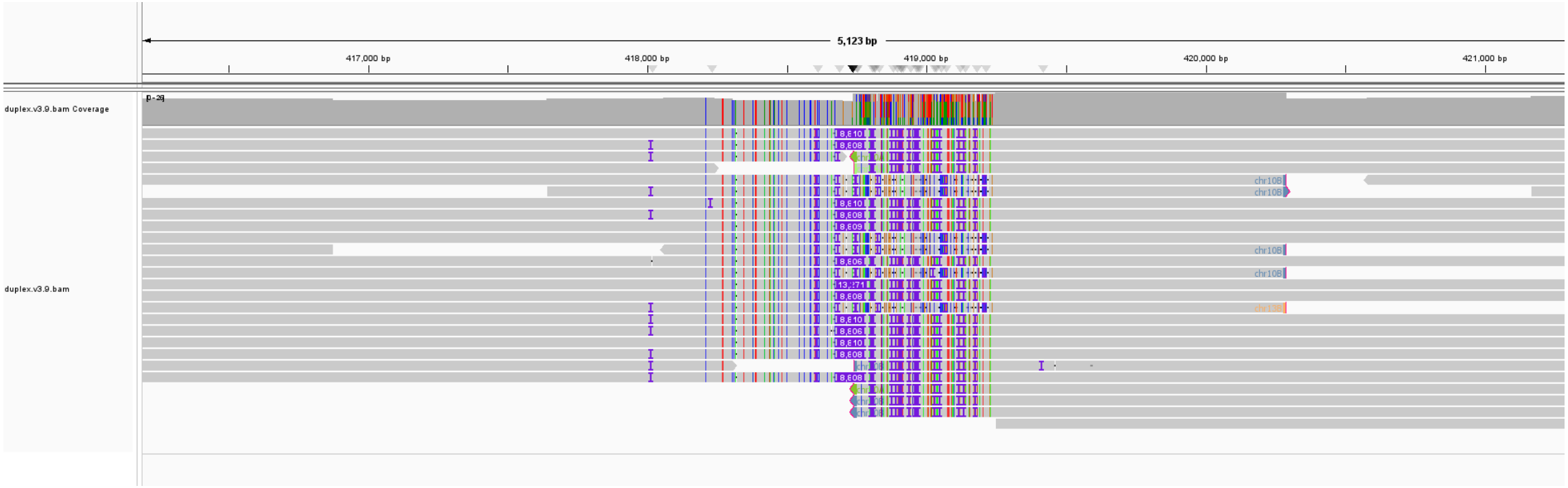
Misjoins

Mapping statistics

k-mer based consensus quality

SNP/indels

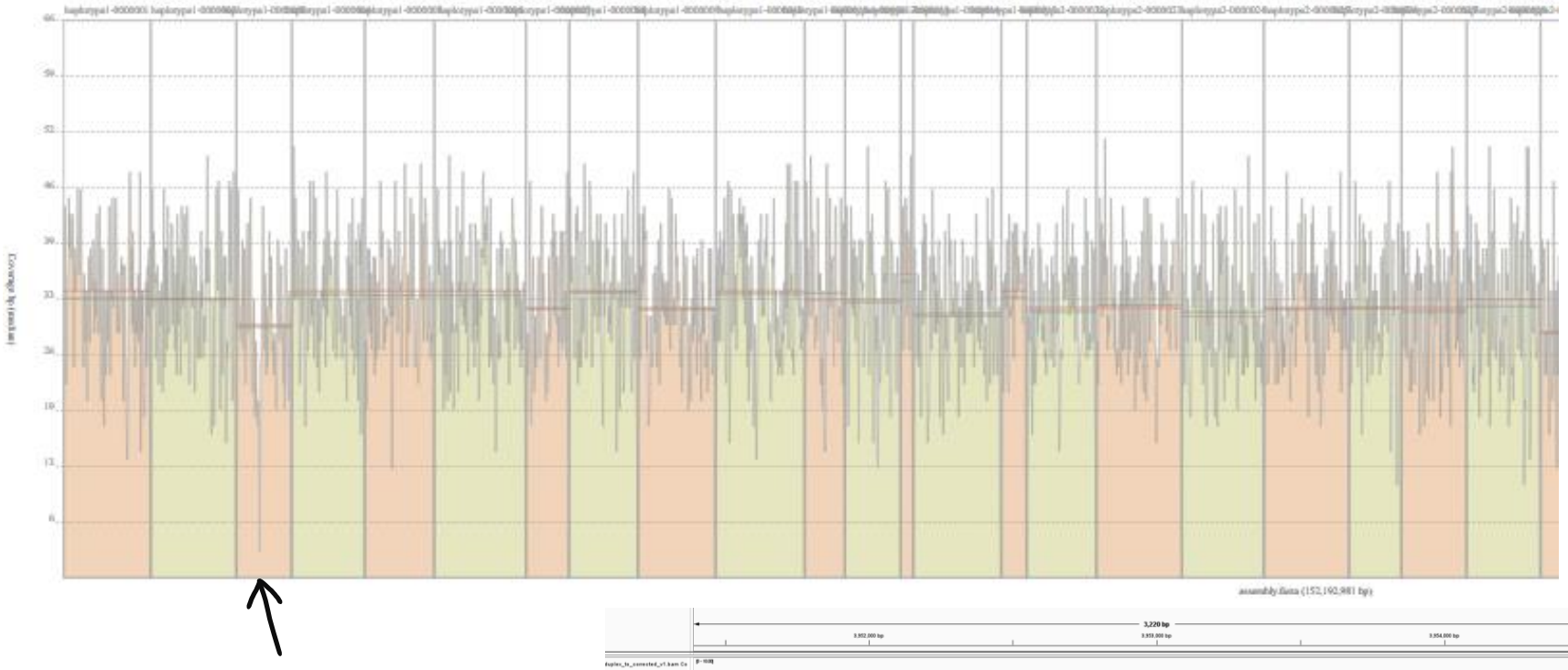
- Map reads back against the assembly
- Ideally should be completely identical
- Spot errors by calling SNP/indels



Structural errors

- Increase in coverage – unresolved repeats/duplications
- Drop in coverage – misjoins, misassemby

When ultralong reads come to rescue



Coverage drop in highly accurate ONT reads

Ultralong ONT reads filled it up!

homopolymer

jvarkit

WGSCoveragePlotter

last commit august

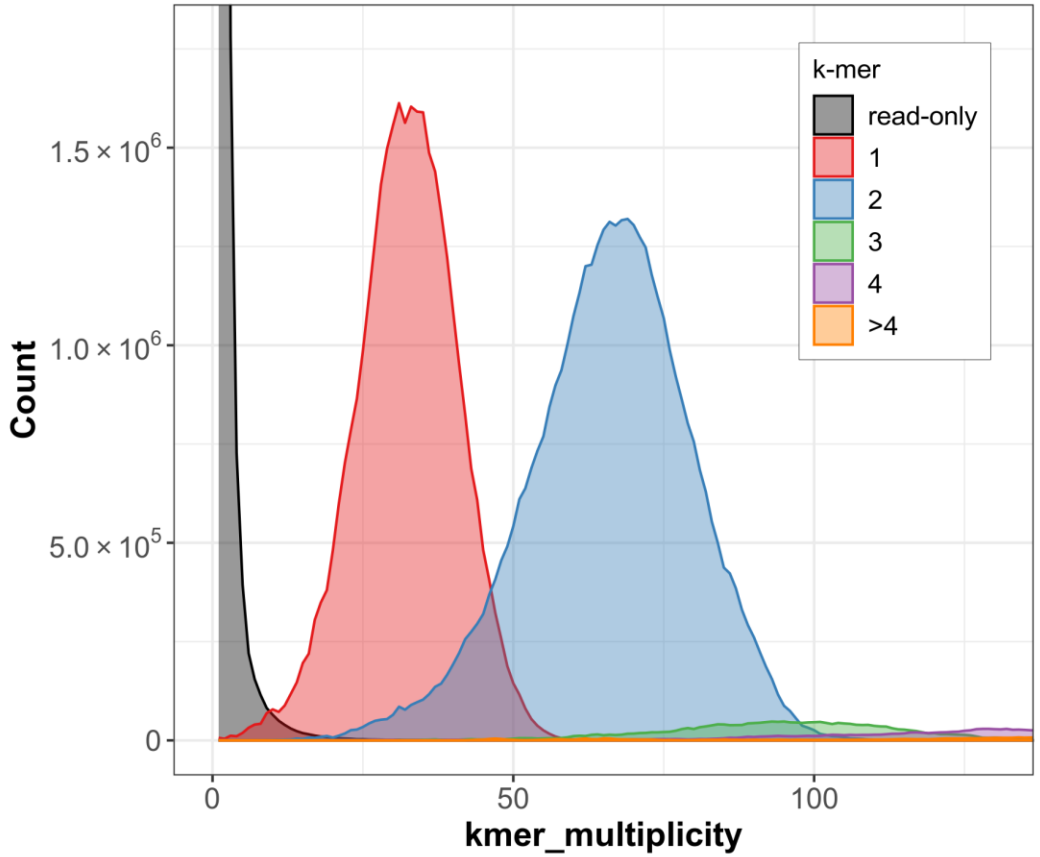
Whole genome coverage plotter

k-mer based assembly evaluation

Merqury

Find concordance between your read dataset and assembly

Reports k-merquality values (QV)




```
v3.9.chr.haplotype-paired       8597    152306908       57.3972 1.82086e-06
duplex_merqury.qv (END)
```

Equates to >99.999% accuracy

Summary

- Never assume your assembly is “error-free”
- Know how trustworthy your assembly is for any analysis

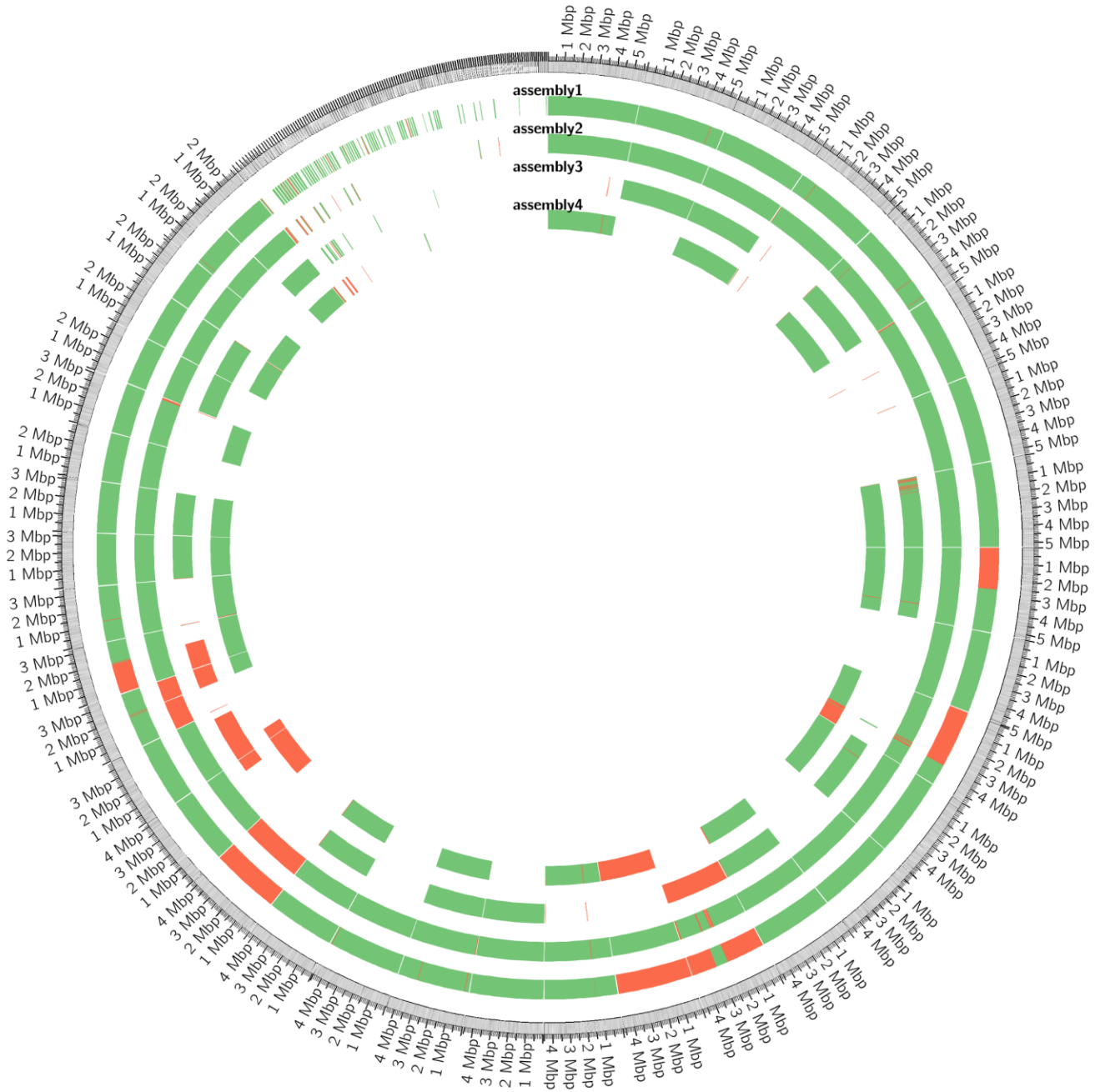
- **The “3C” rules**
  - Contiguity
  - Completeness
  - Correctness

“Is my assembly ready for scaffolding?”

“Is my assembly ready for gene annotation?”

Wanna compare multiple assemblies and choose the best one?

Try QUAST circo plot!



Different read data types + assembler tool

1: ONT reads + Verkko

2: error-corrected ONT reads + Verkko

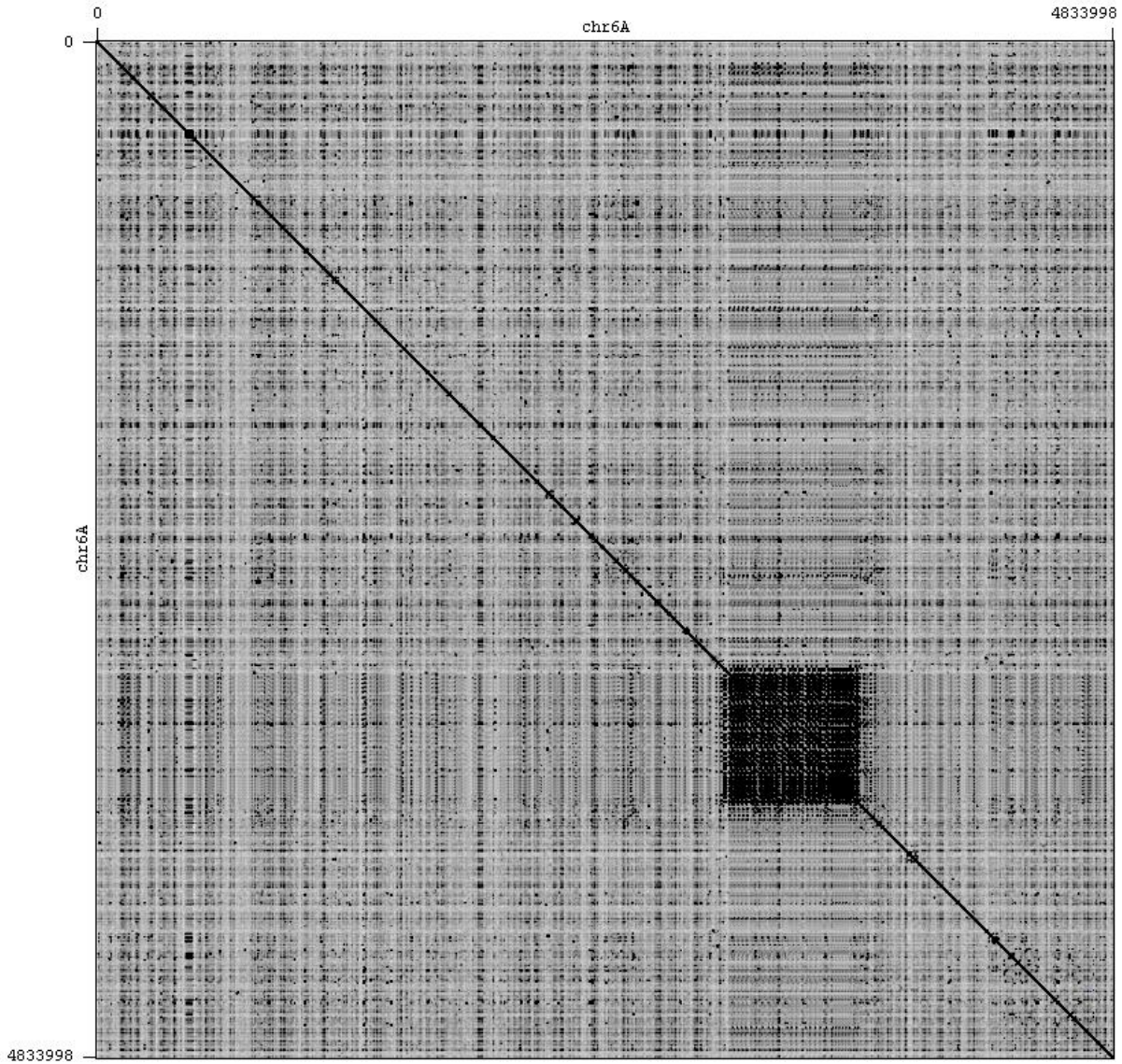
3: ONT reads + Hifiasm (failed run)

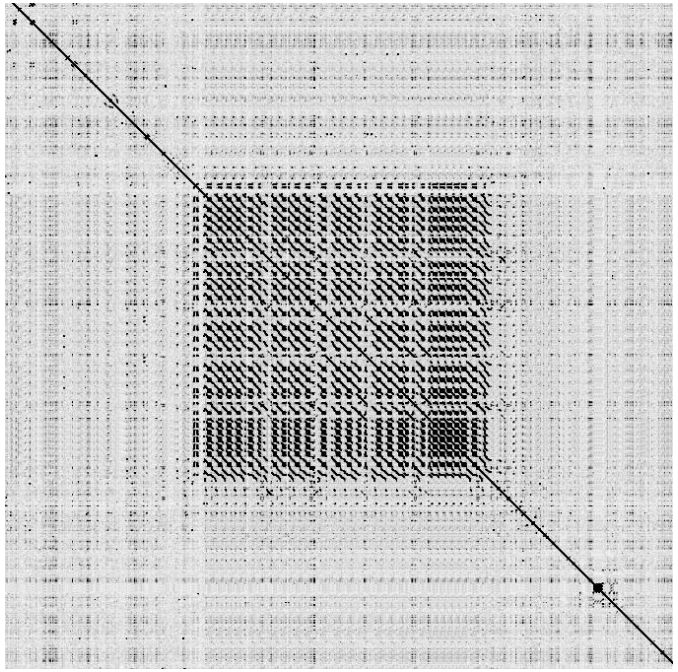
4: error-corrected ONT reads + Hifiasm (failed run)

Julian Rodriguez Algaba

Self alignment dotplot to find repetitive regions

Gepard, Chromeister





Last important tip

Keep track of your assembly versions as you curate it!

assembly_version_history.txt

```
raw: assembly.raw.fasta

v3.1: assembly.extend_telomere.mtDNA_lowcov_rm.fasta

v3.2: telomeres extended. telomeric motif copy number corrected for all
chromosomes.
v3.2.1: further telomere correction of v3.2 after coverage check.
extension stops when last supported by at least two reads (2 simplex, or 1
simplex 1 duplex.)

v3.3: after HiC scaffold. (v3.2.1.FINAL.fasta output from 3d-dna)
orientation swapped to make sure p arm starts first.

v3.4: replace chr5A with v2.6 manually scaffolded chr5A. telomeres
extended.

v3.5: fixed chr17B gap associated with a ~500bp GAAAA tandem repeat.

v3.5.1: fixed misassembly before the 5kbp rDNA-associated gap Ns. (Flye to
assemble HiC_scaffold_44,45,46 into a high quality local assembly. mapped
back to chr17B and replaced 89647bp before gap.)
```