

Genome assembly evaluation: the good, the bad, and the ugly

Nadège Guiglielmoni

Sequencing data analysis

Check quality/content, remove adaptors

- ▶ **Illumina**

- ▶ quality control: fastqc
- ▶ adapter removal: cutadapt, Trimmomatic, Sequencher...
- ▶ *k*-mer analysis: GenomeScope, KAT, smudgeplot

- ▶ **Nanopore**

- ▶ quality control: Nanoplot
- ▶ adapter removal: Nanopore tools, Porechop

- ▶ **HiFi**

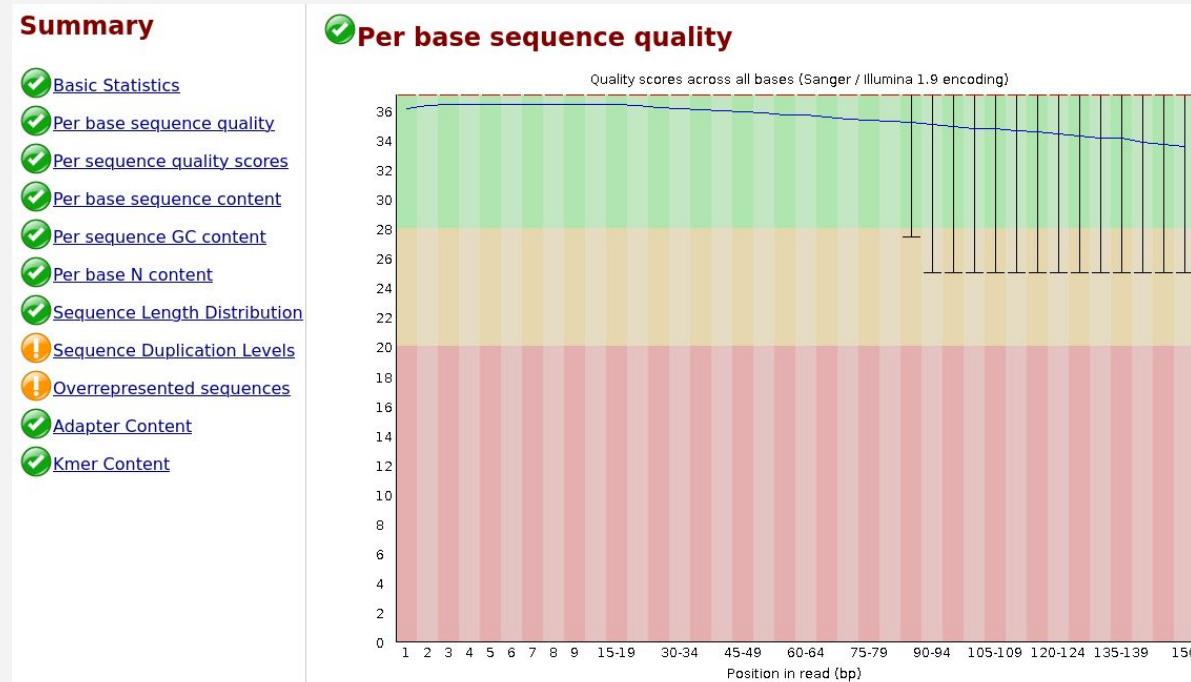
- ▶ quality control: Nanoplot
- ▶ *k*-mer analysis: GenomeScope, KAT, smudgeplot

Sequencing data analysis

- ▶ **Illumina**
 - ▷ quality control: fastqc
- ▶ **Nanopore**
 - ▷ quality control: Nanoplot
- ▶ **HiFi**
 - ▷ quality control: Nanoplot

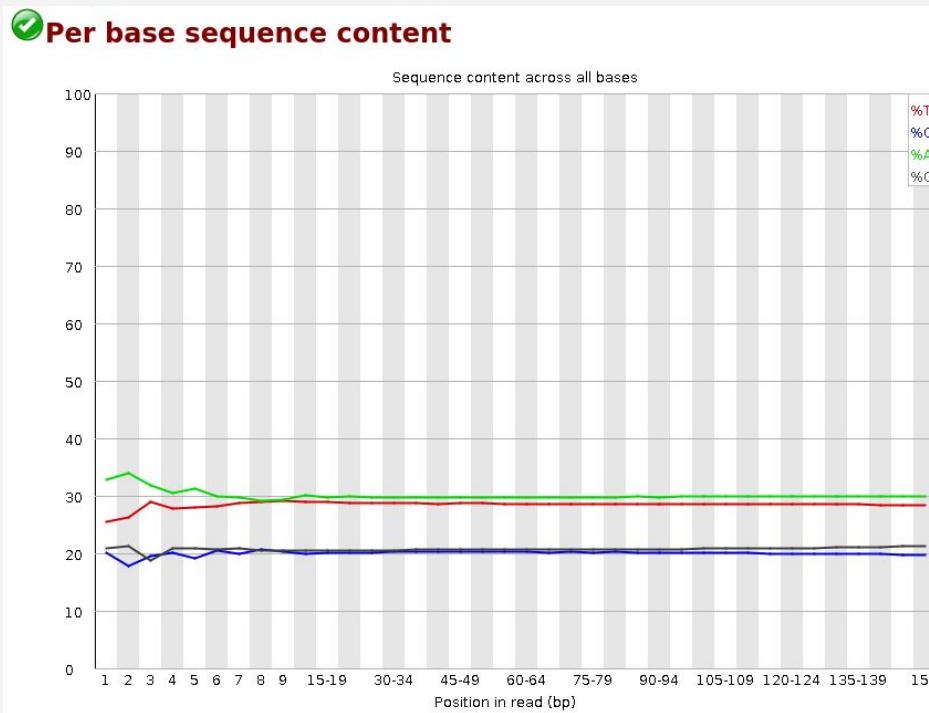
Sequencing data analysis

Illumina fastqc



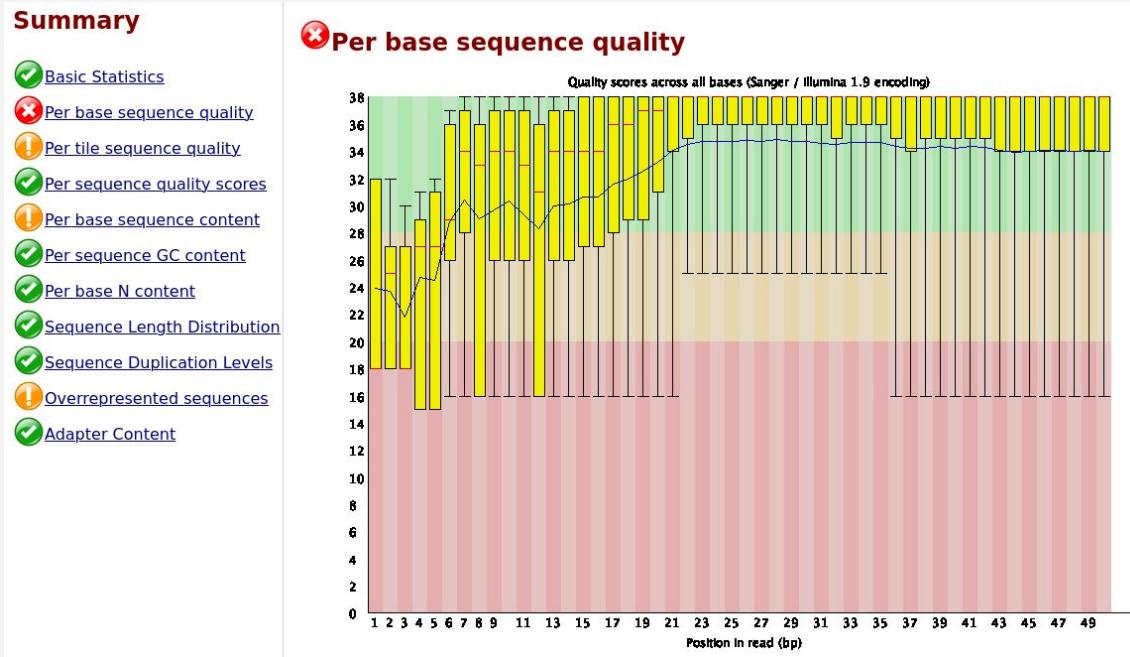
Sequencing data analysis

Illumina fastqc



Sequencing data analysis

Illumina fastqc

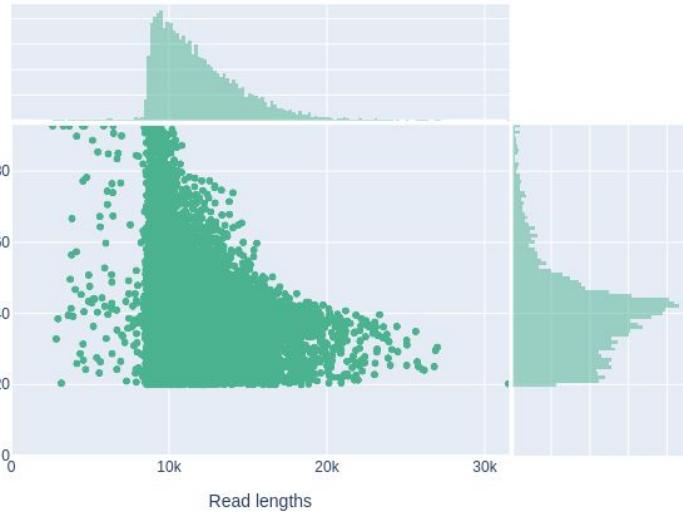


Sequencing data analysis

NanoPlot

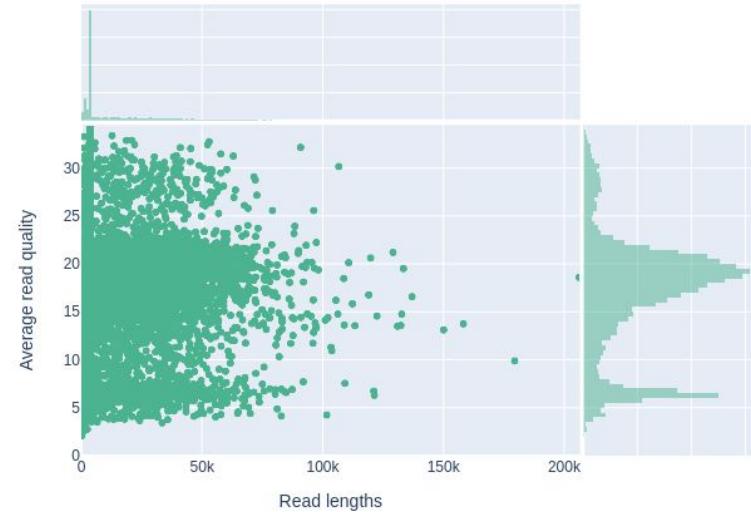
PacBio HiFi

Read lengths vs Average read quality plot using dots



Nanopore

Read lengths vs Average read quality plot using dots

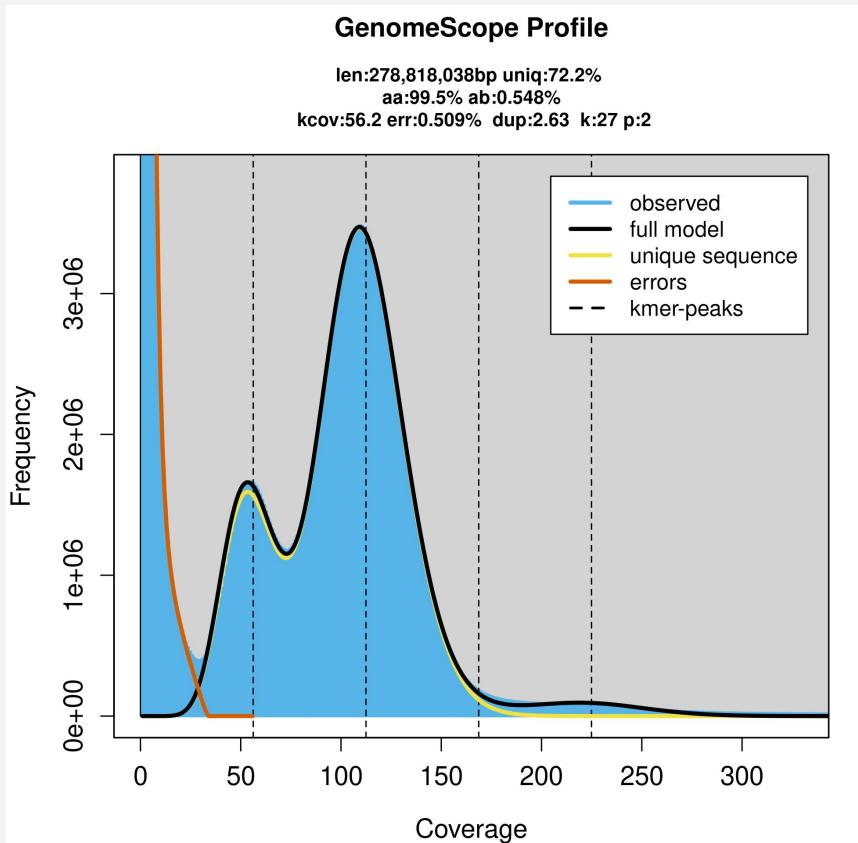


Sequencing data analysis

- ▶ Illumina
- ▶ HiFi
 - ▷ *k*-mer analysis: GenomeScope, KAT, Smudgeplot

Sequencing data analysis

Gordionus montsenyensis

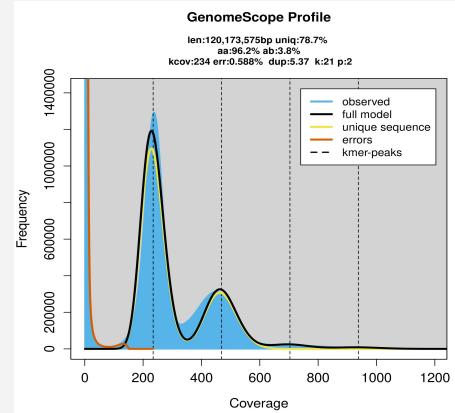


The genome sequence of the Montseny horsehair worm, *Gordionus montsenyensis* sp. nov., a key resource to investigate Ecdysozoa evolution.
Eleftheriadi et al. bioRxiv, 2023.

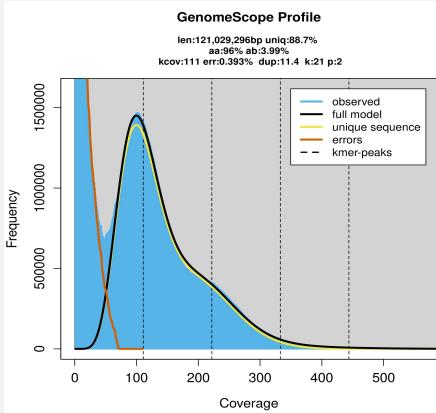
Sequencing data analysis

Plectus sambesii

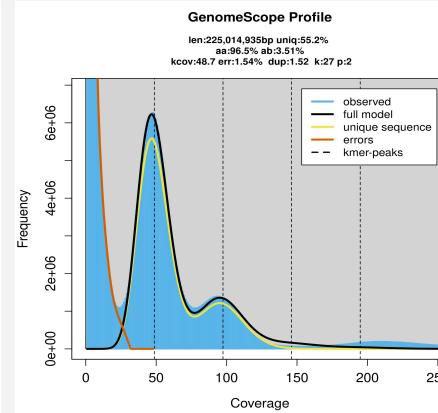
Illumina



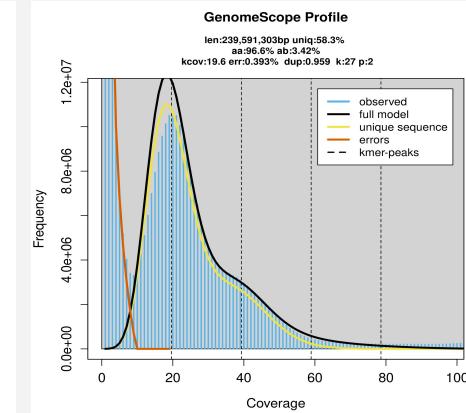
PacBio HiFi



Nanopore

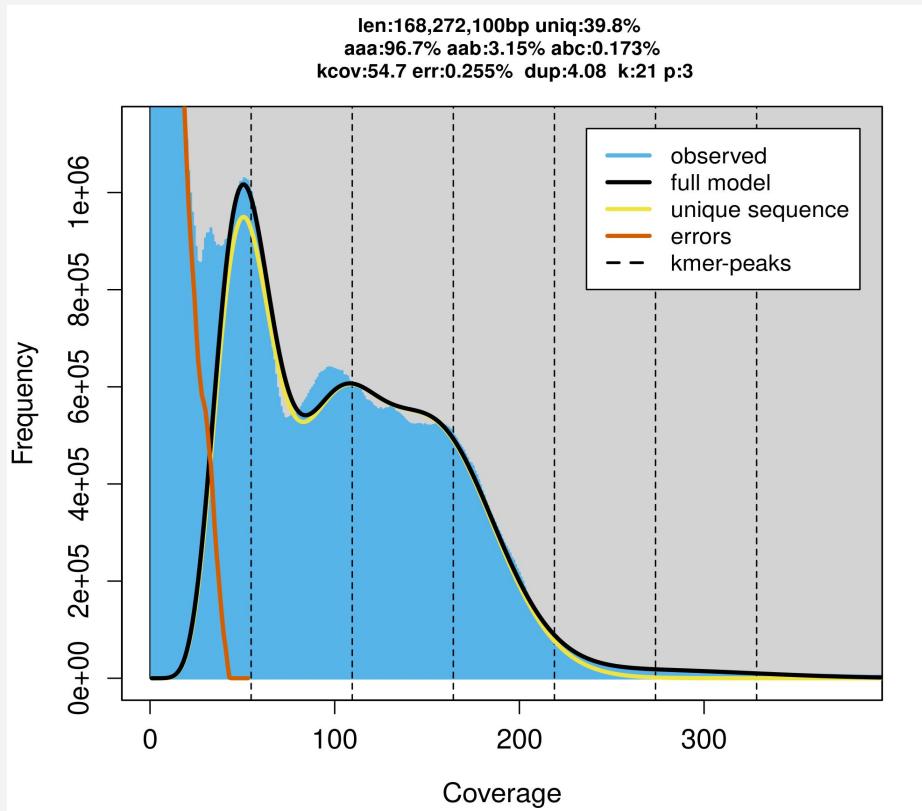


Nanopore Q20



Sequencing data analysis

Panagrolaimus PS1159

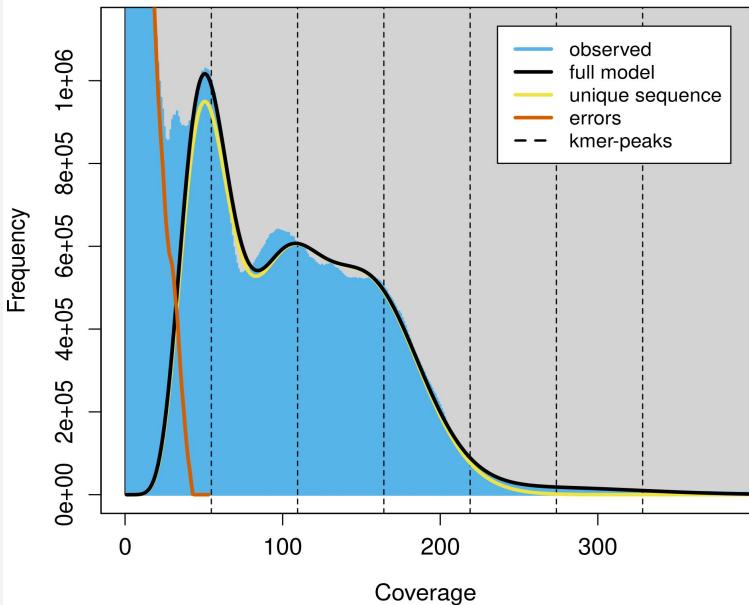


Sequencing data analysis

Panagrolaimus PS1159

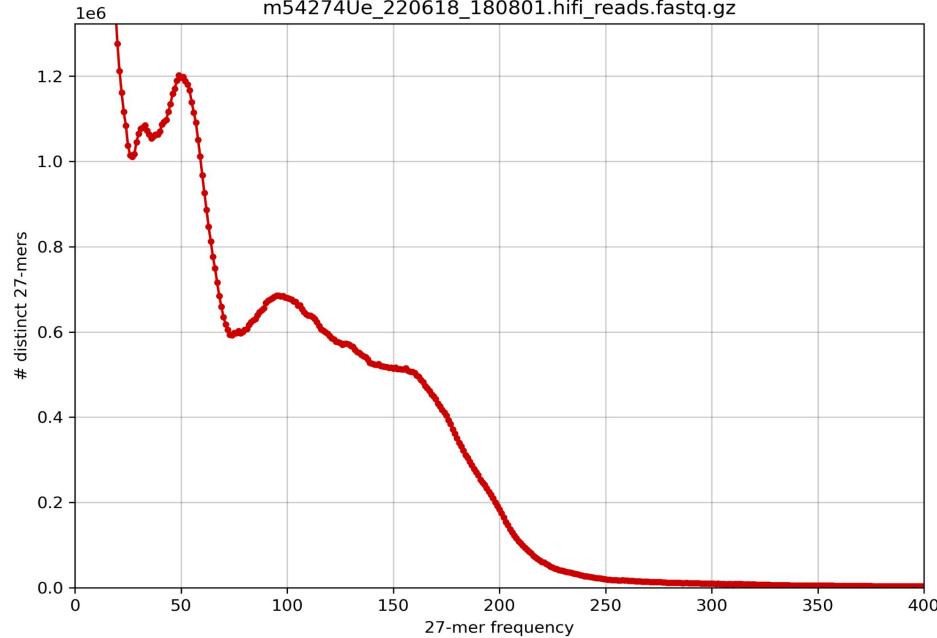
GenomeScope

len:168,272,100bp uniq:39.8%
aaa:96.7% aab:3.15% abc:0.173%
kcov:54.7 err:0.255% dup:4.08 k:21 p:3



KAT hist

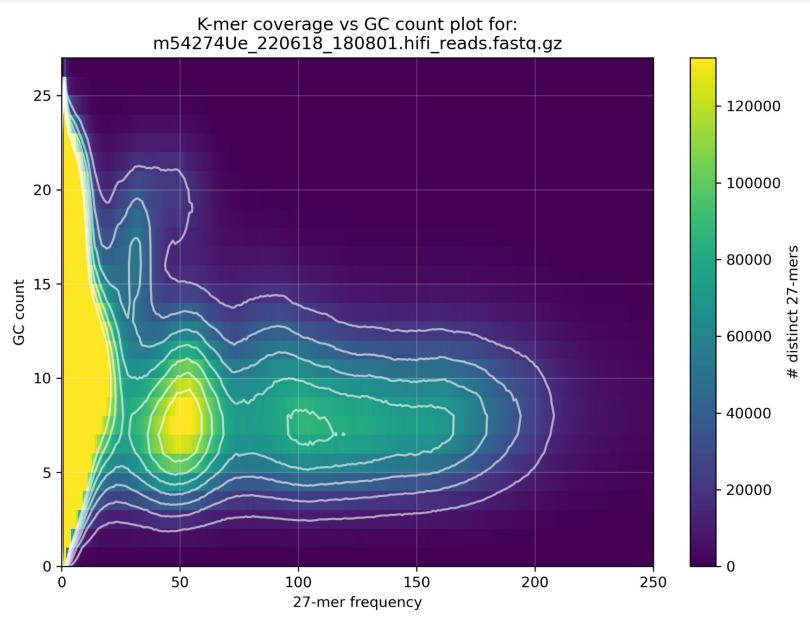
27-mer spectra for:
m54274Ue_220618_180801.hifi_reads.fastq.gz



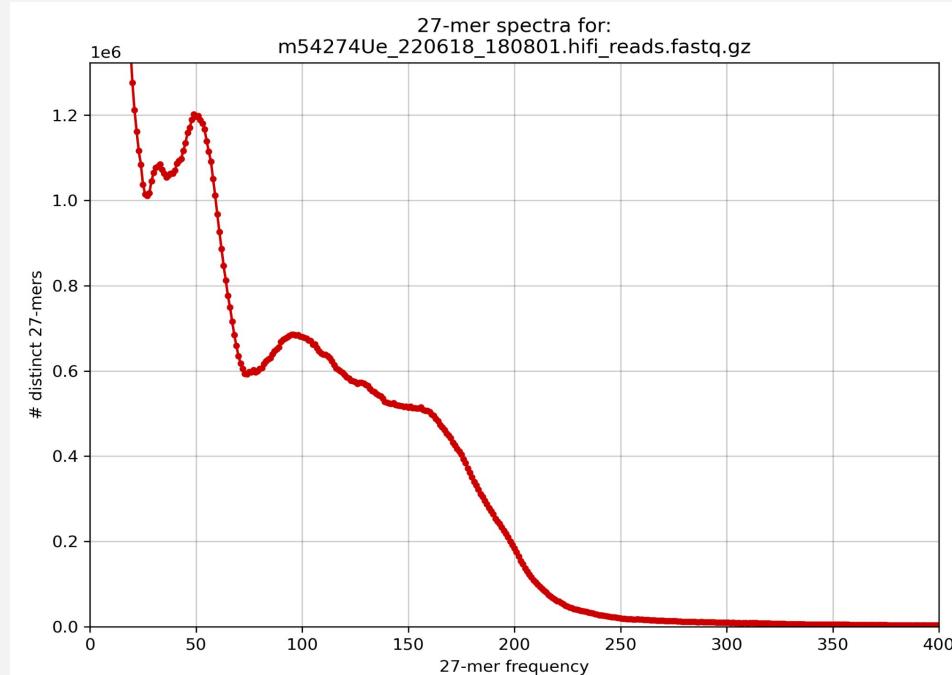
Sequencing data analysis

Panagrolaimus PS1159

KAT gcp



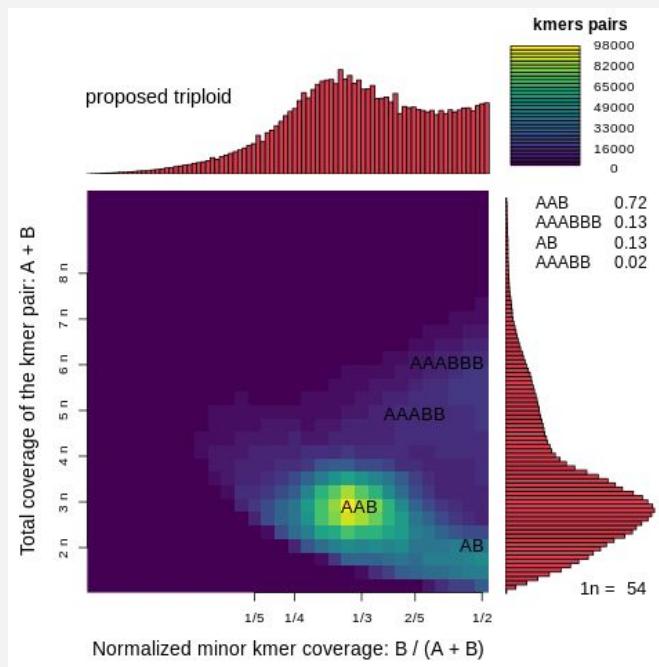
KAT hist



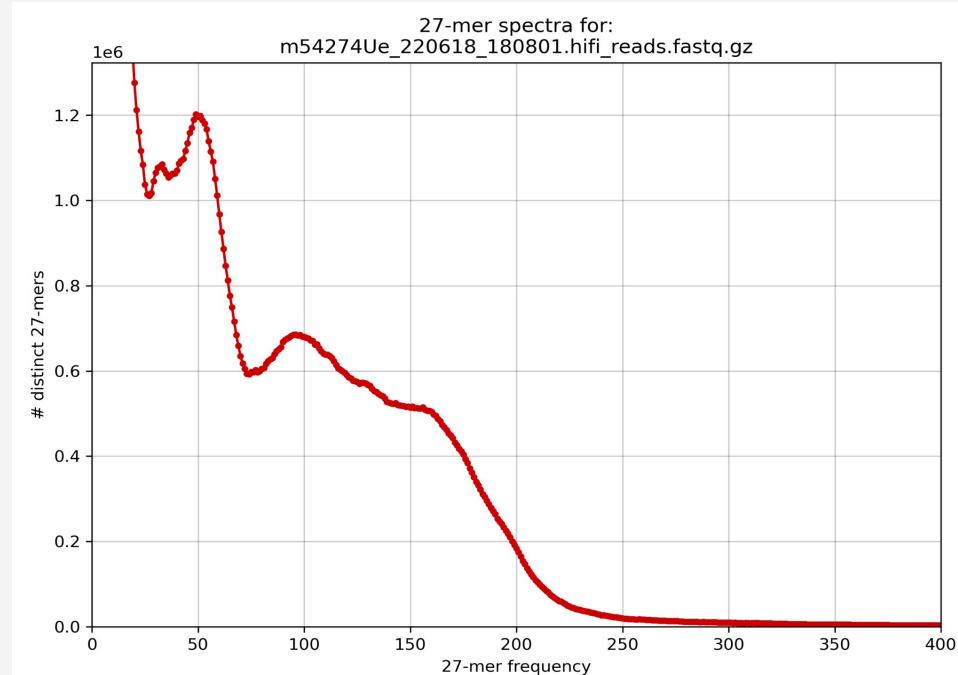
Sequencing data analysis

Panagrolaimus PS1159

Smudgeplot

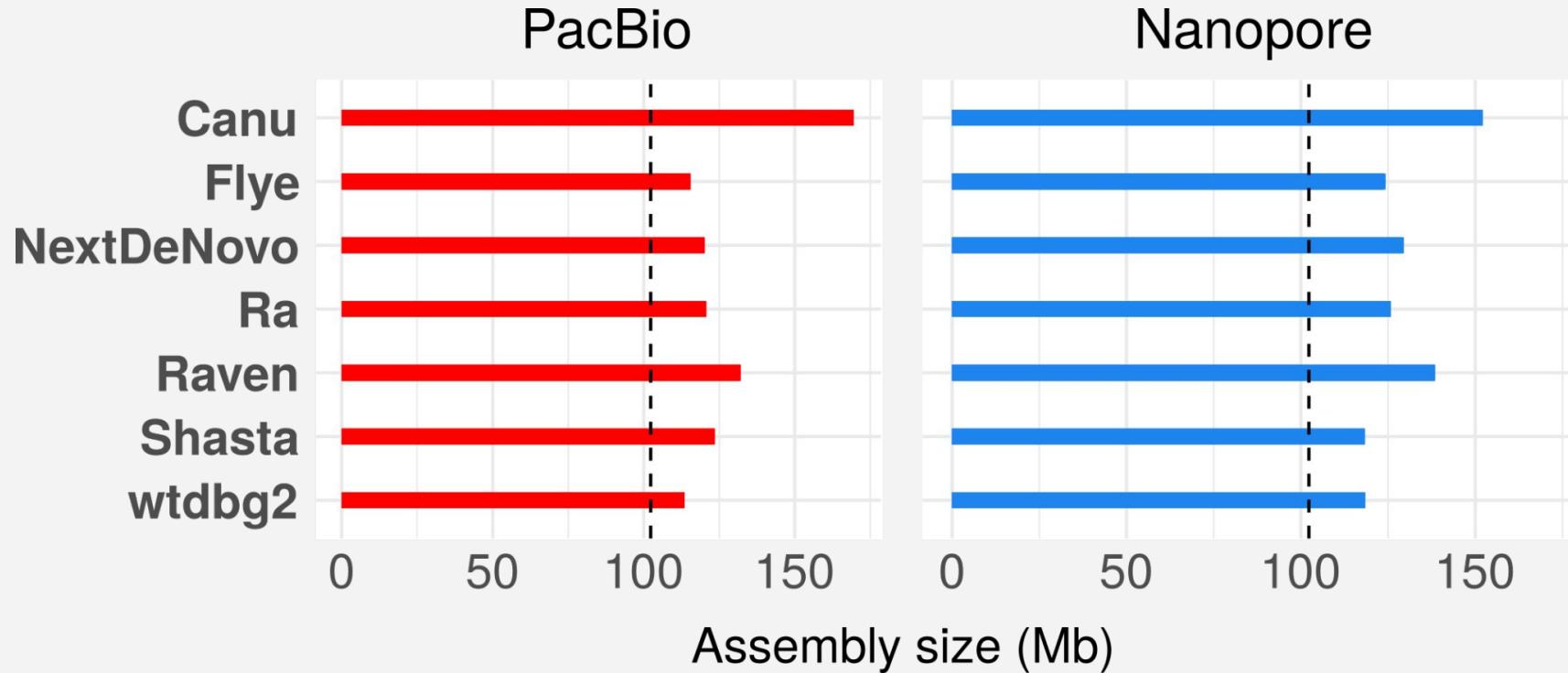


KAT hist



Assembly size

Adineta vaga



Contiguity

The **length** of the largest contig (or scaffold) for which **50% of the assembly size** is contained in contigs (or scaffolds) of **equal or greater length**

Contiguity

The **length** of the largest contig (or scaffold) for which **50% of the assembly size** is contained in contigs (or scaffolds) of **equal or greater length**



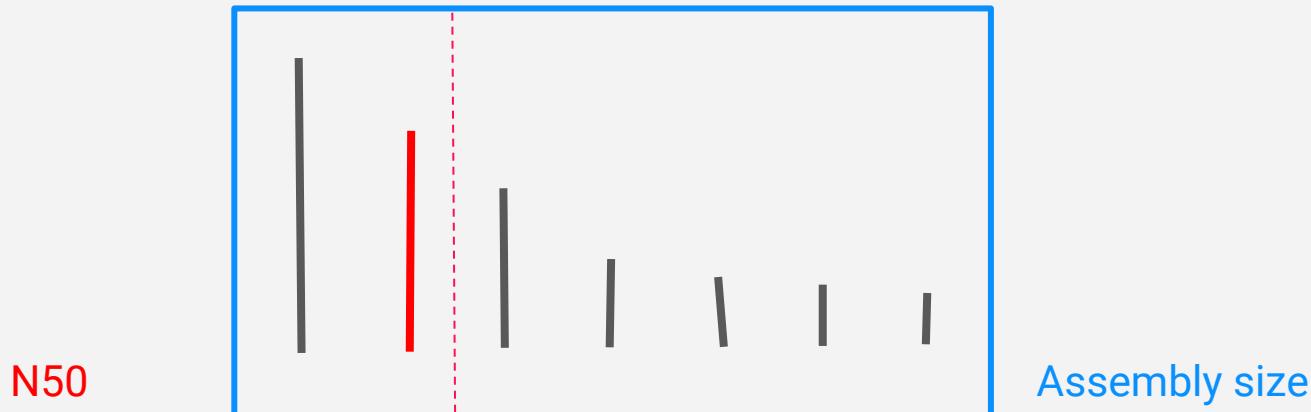
Contiguity

The **length** of the largest contig (or scaffold) for which **50% of the assembly size** is contained in contigs (or scaffolds) of **equal or greater length**



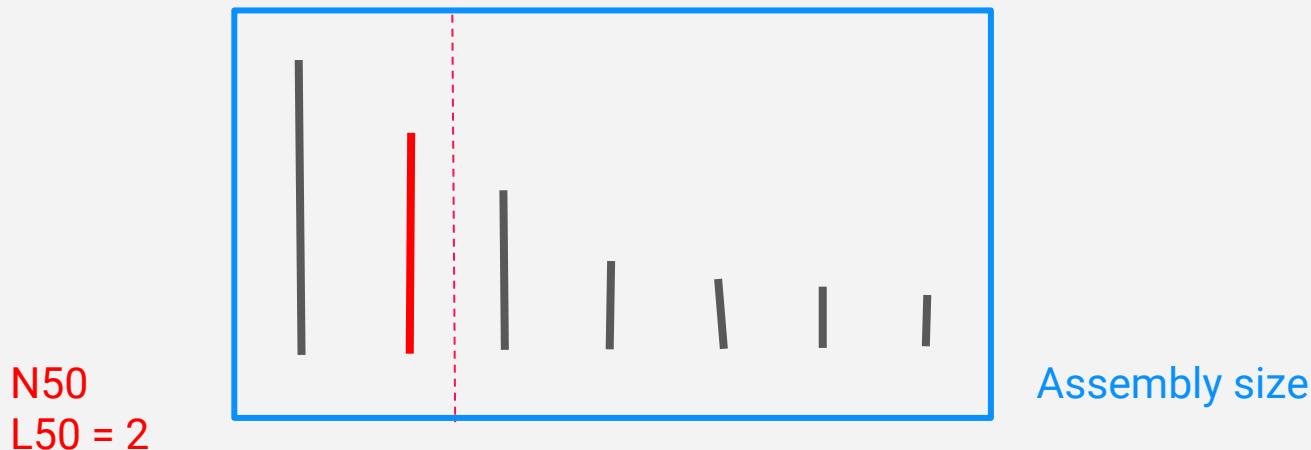
Contiguity

The **length** of the largest contig (or scaffold) for which **50% of the assembly size** is contained in contigs (or scaffolds) of **equal or greater length**



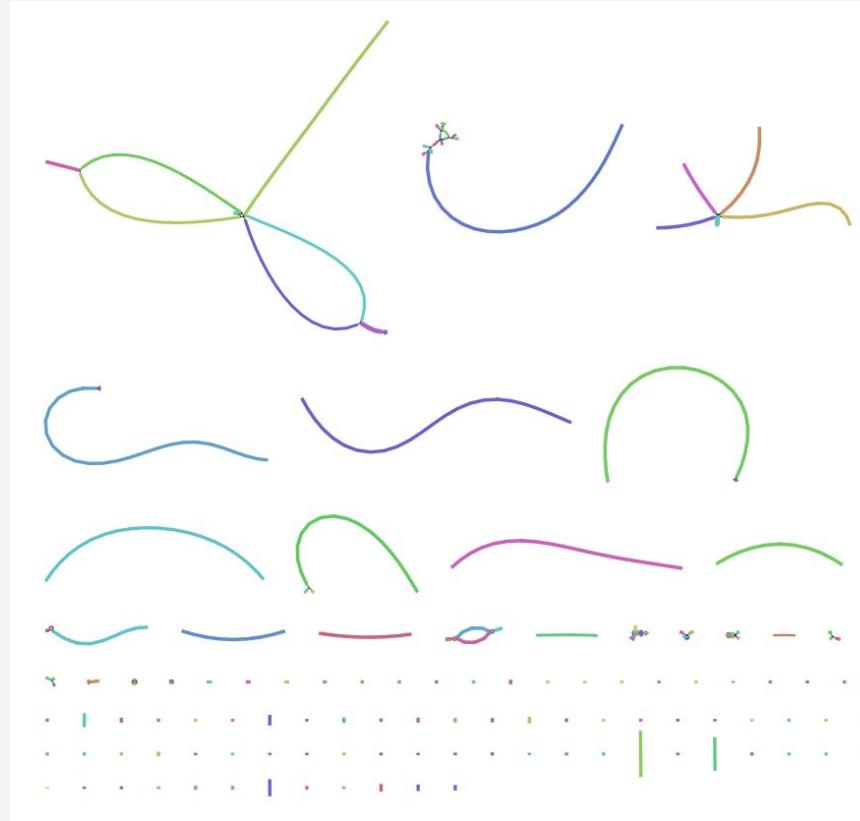
Contiguity

The **length** of the largest contig (or scaffold) for which **50% of the assembly size** is contained in contigs (or scaffolds) of **equal or greater length**



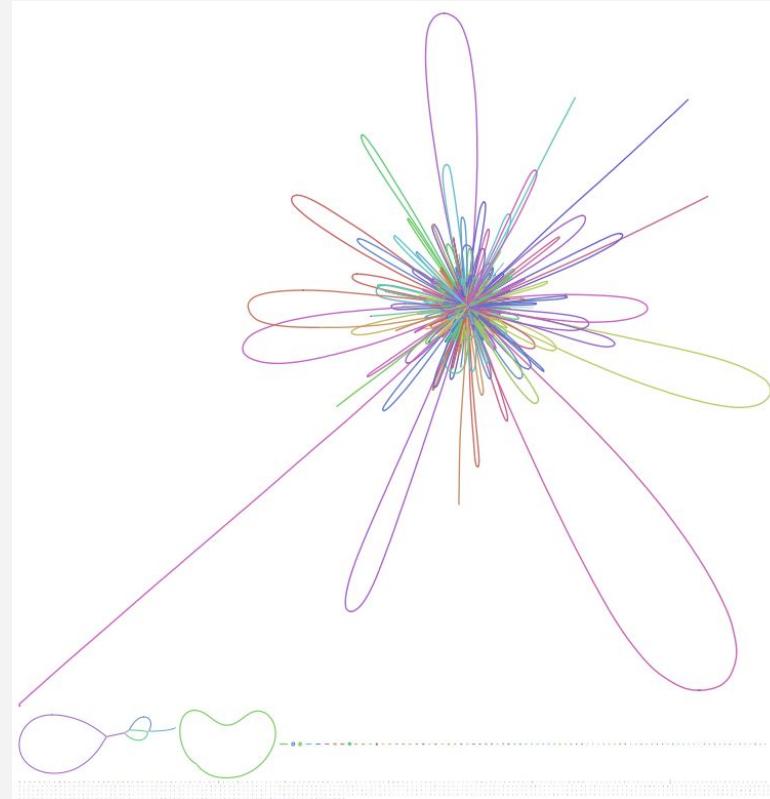
Assembly graph

Bandage



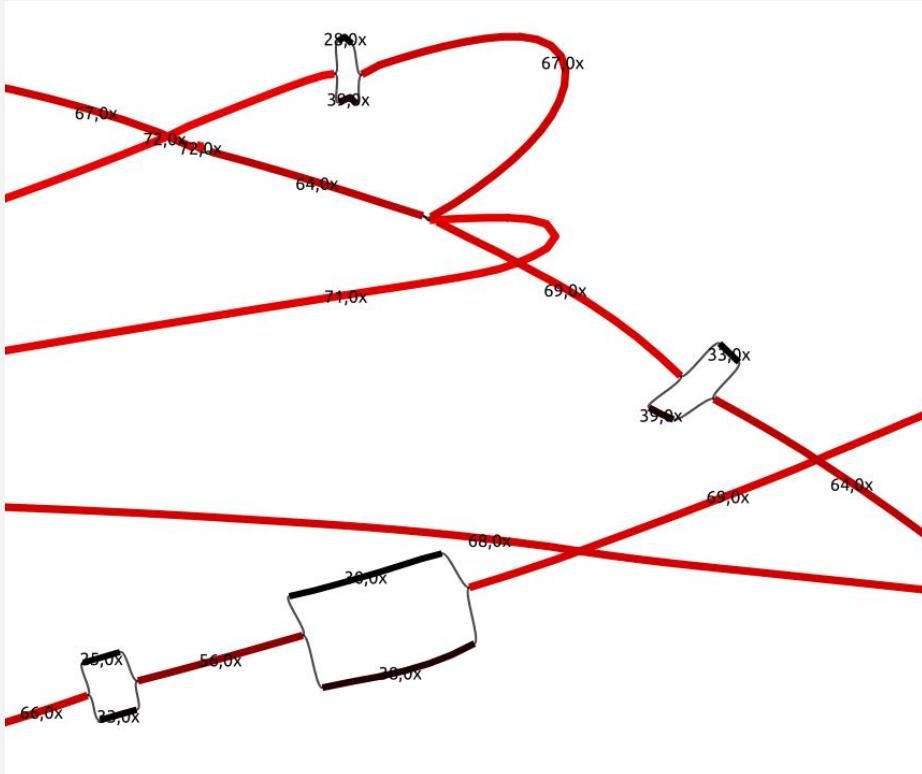
Assembly graph

Bandage



Assembly graph

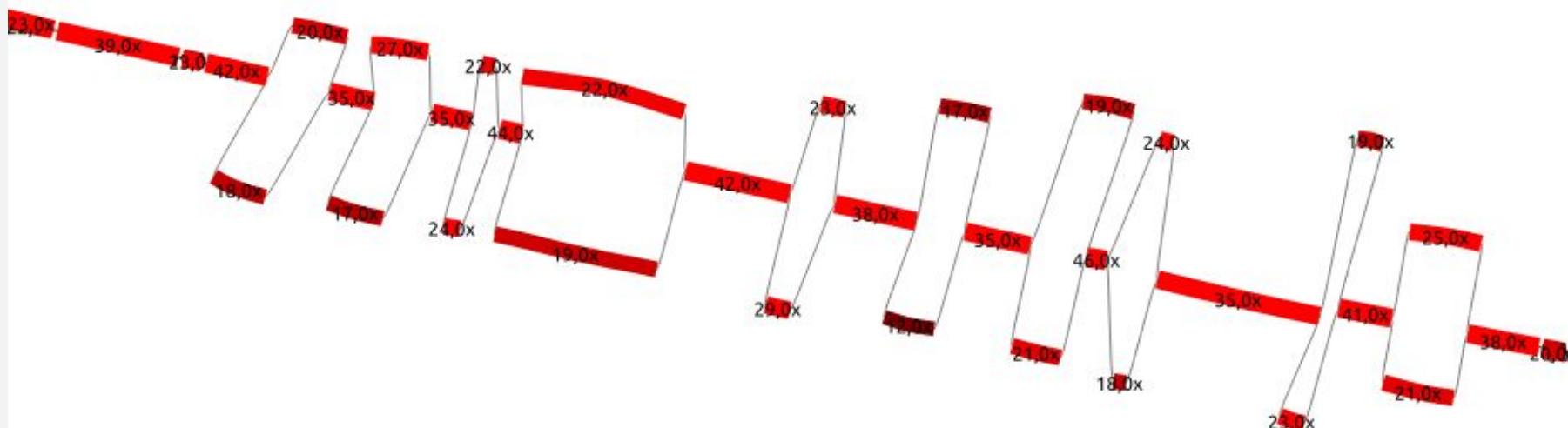
Bandage



bubbles

Assembly graph

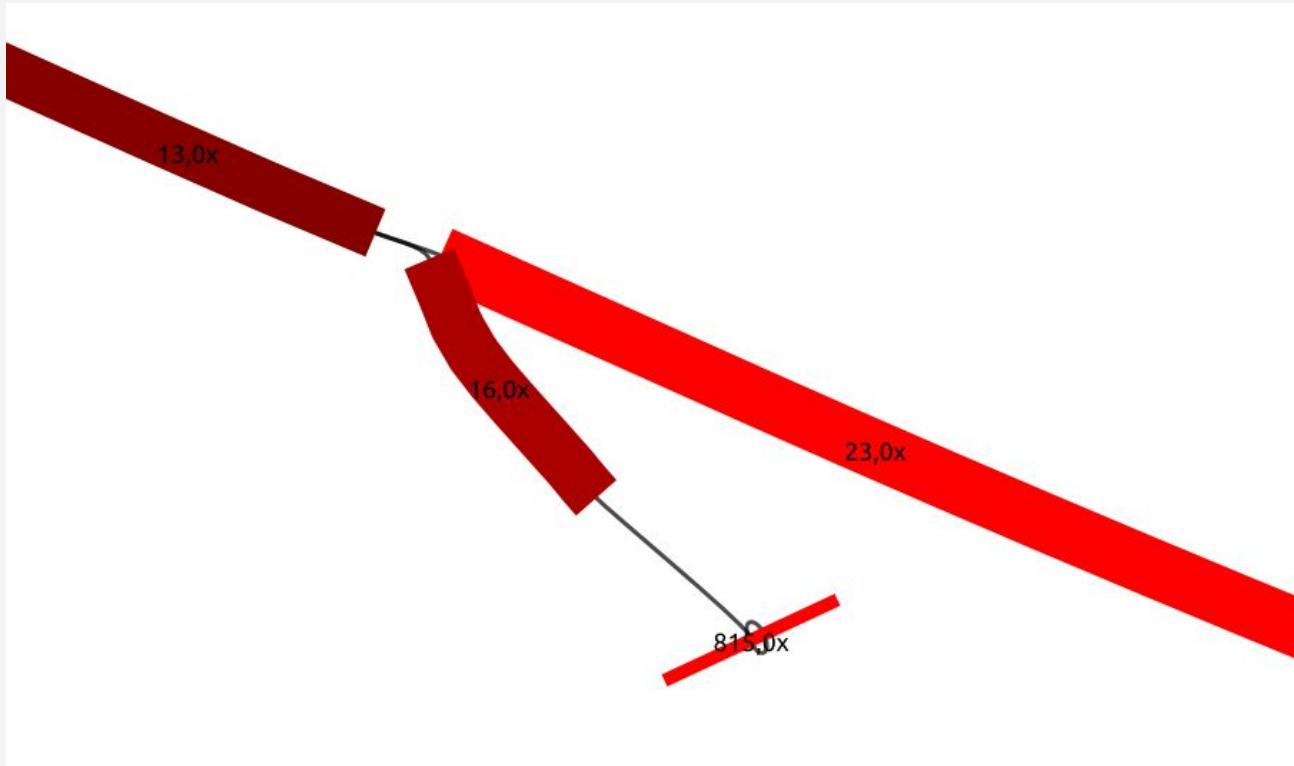
Bandage



bubbles

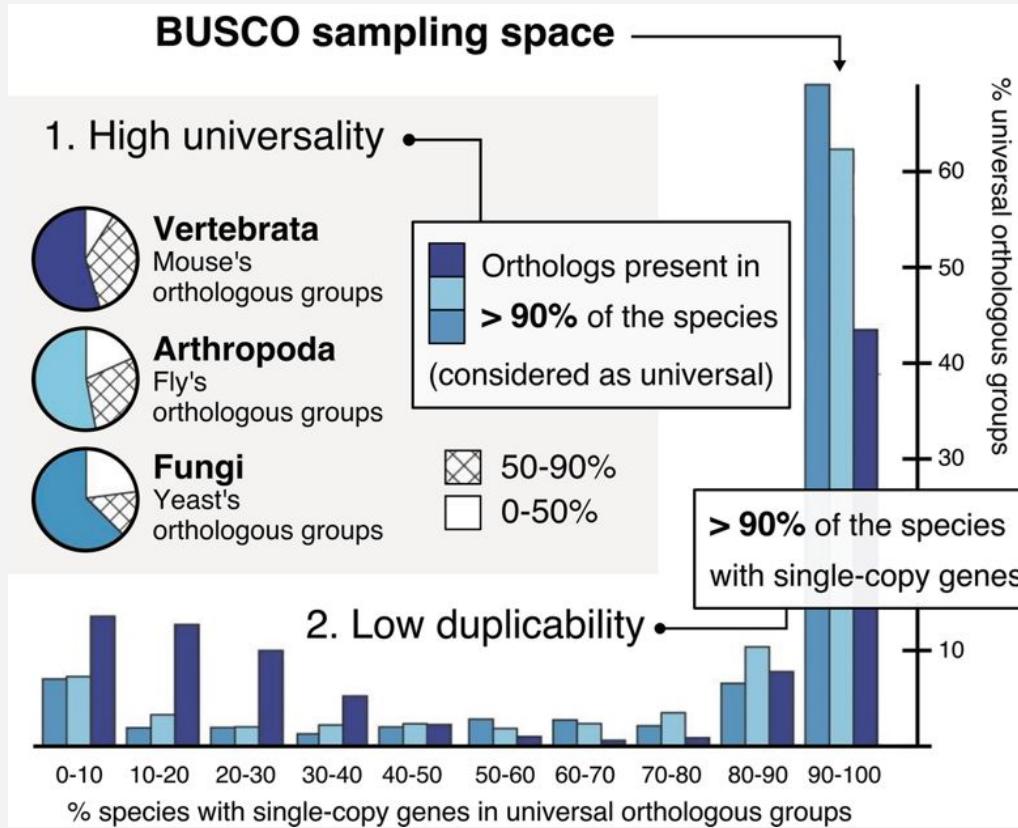
Assembly graph

Bandage



repetitions

Ortholog completeness



Ortholog completeness

Adineta vaga

Reads	Assembly	Complete features	Complete single-copy	Complete duplicated	Fragmented	Missing
Illumina	BWISE	82.6%	16.4%	66.2%	3.2%	14.2%
PacBio	NextDenovo	84.1%	72.0%	12.1%	2.8%	13.1%
Nanopore	NextDenovo	44.7%	39.6%	5.1%	21.1%	34.2%
Nanopore + Illumina	NextDenovo	86.2%	72.0%	14.2%	1.9%	11.9%
HiFi	hifiasm	85.8%	18.7%	67.1%	1.8%	12.4%

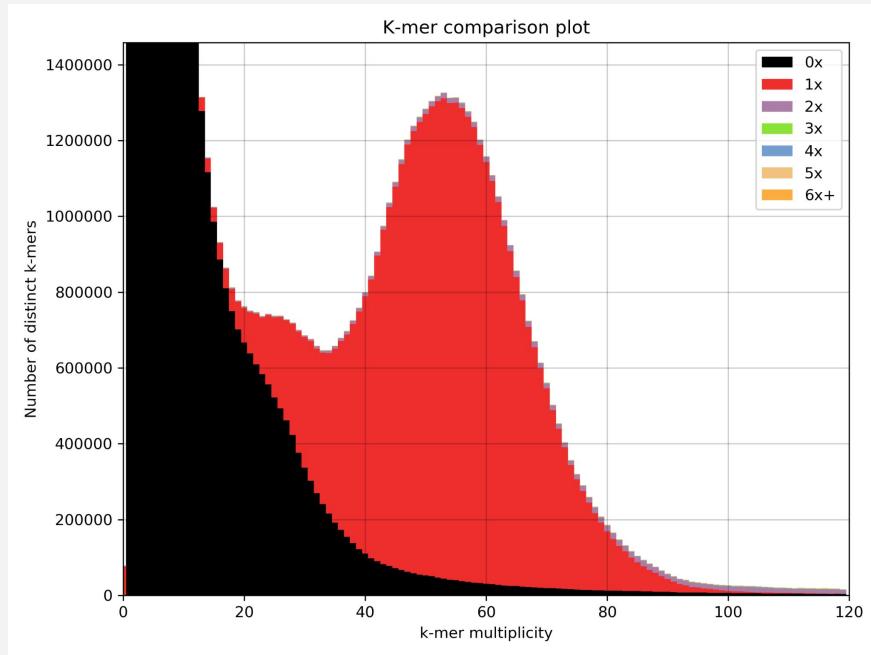
Ortholog completeness

Flaccisagitta enflata

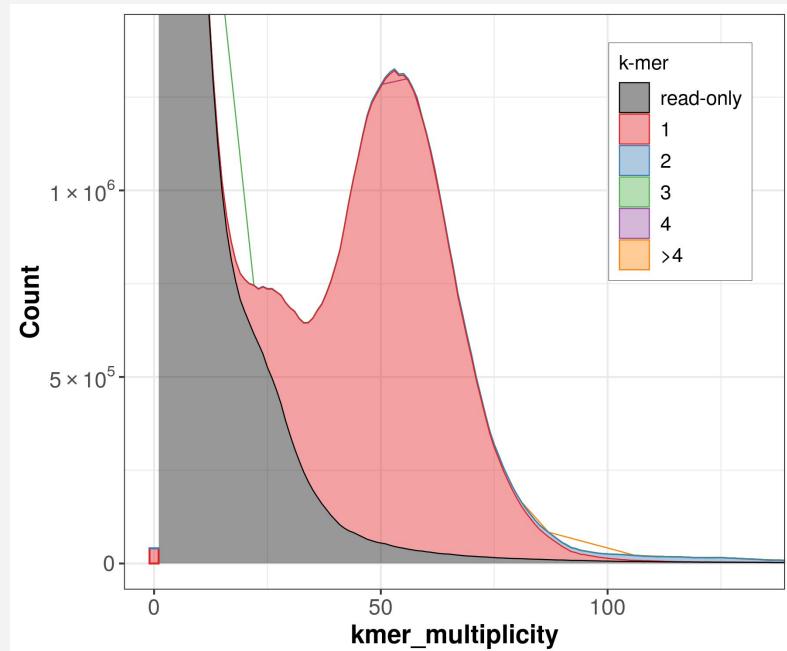
Assembly	Assembly size	N50	Single-copy BUSCOs	Duplicated BUSCOs
Ratatosk + Canu	1.48 Gb	195 kb	28.3%	66.0%
Ratatosk + Canu + purge_dups x2	946 Mb	215 kb	85.4%	7.1%
Final scaffolds	794 Mb	91.2 Mb	86.9%	6.4%

k-mer completeness

KAT comp



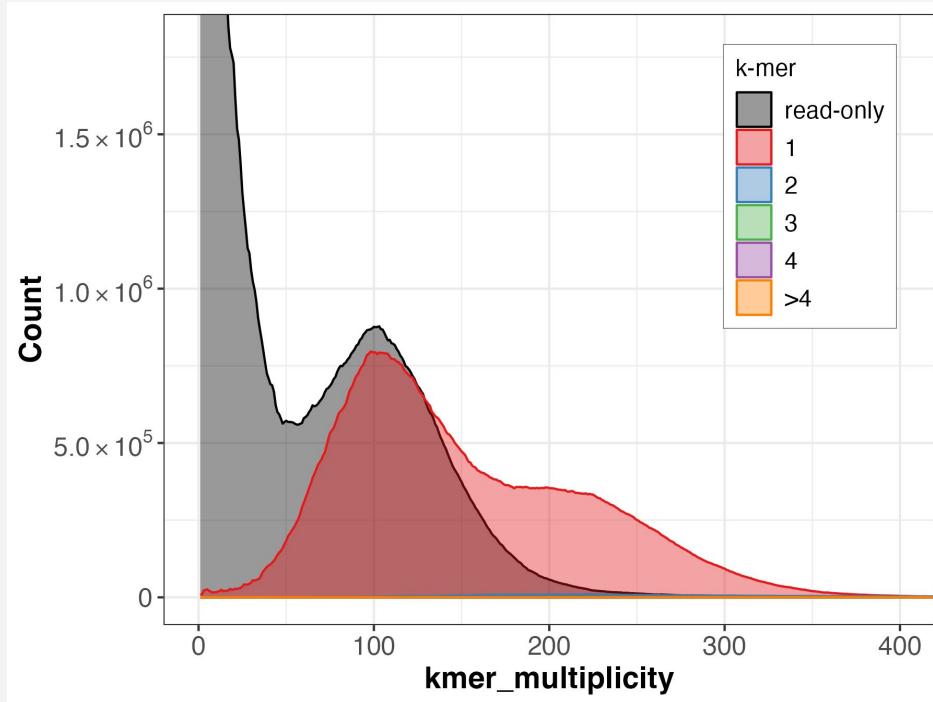
Merqury



k-mer completeness

Plectus sambesii

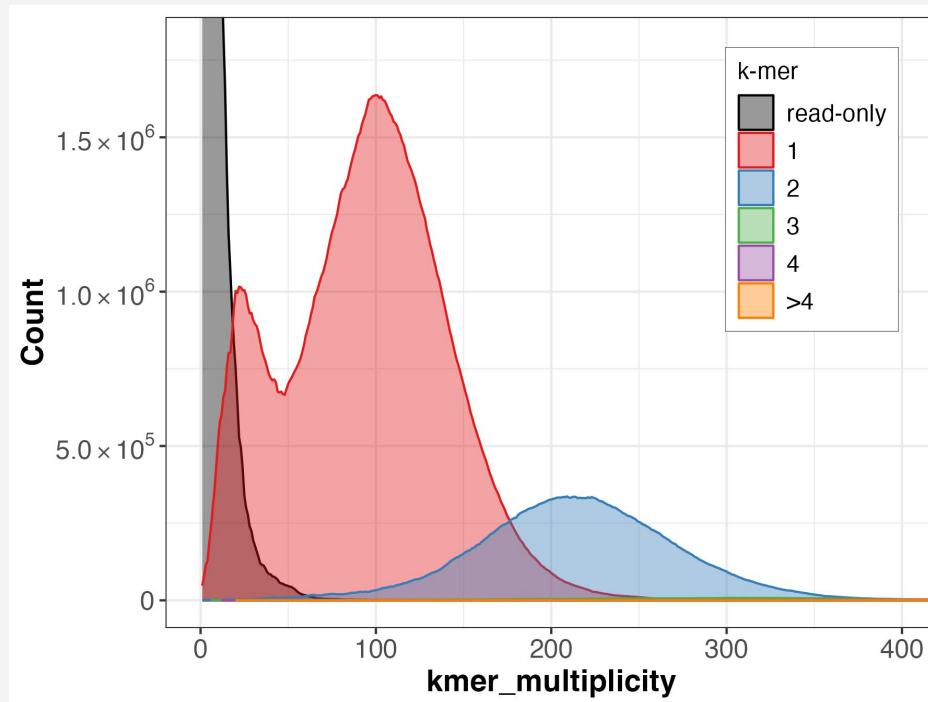
Merqury



k-mer completeness

Plectus sambesii

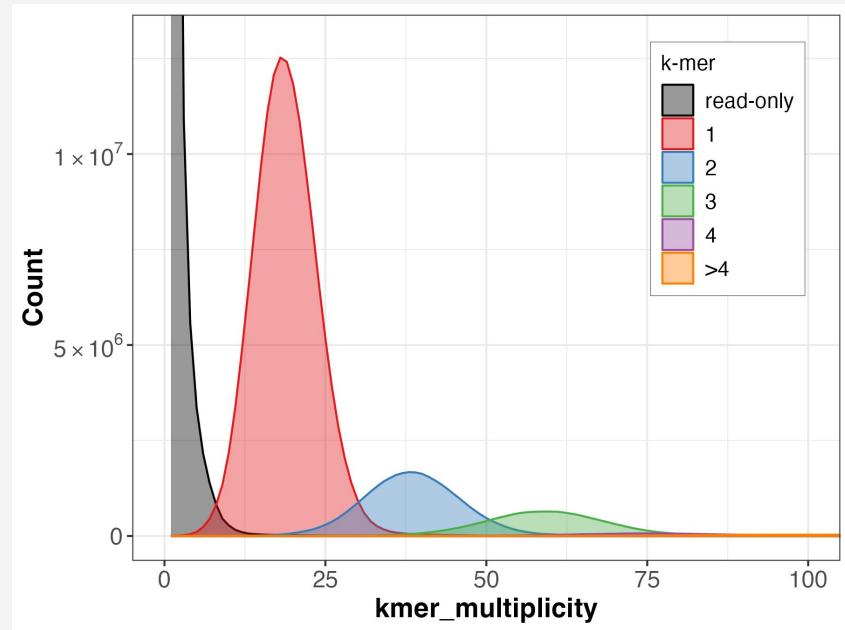
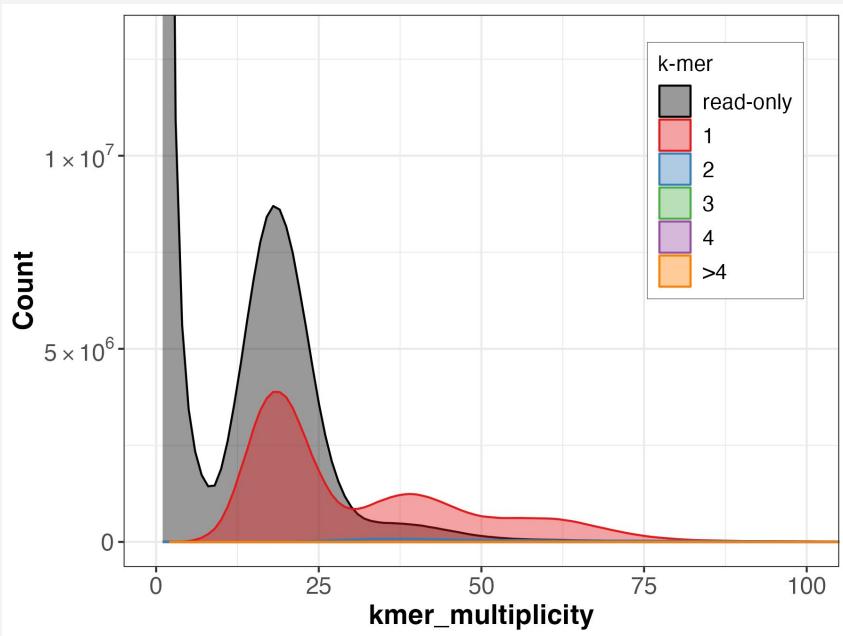
Merqury



k-mer completeness

Panagrolaimus PS1579

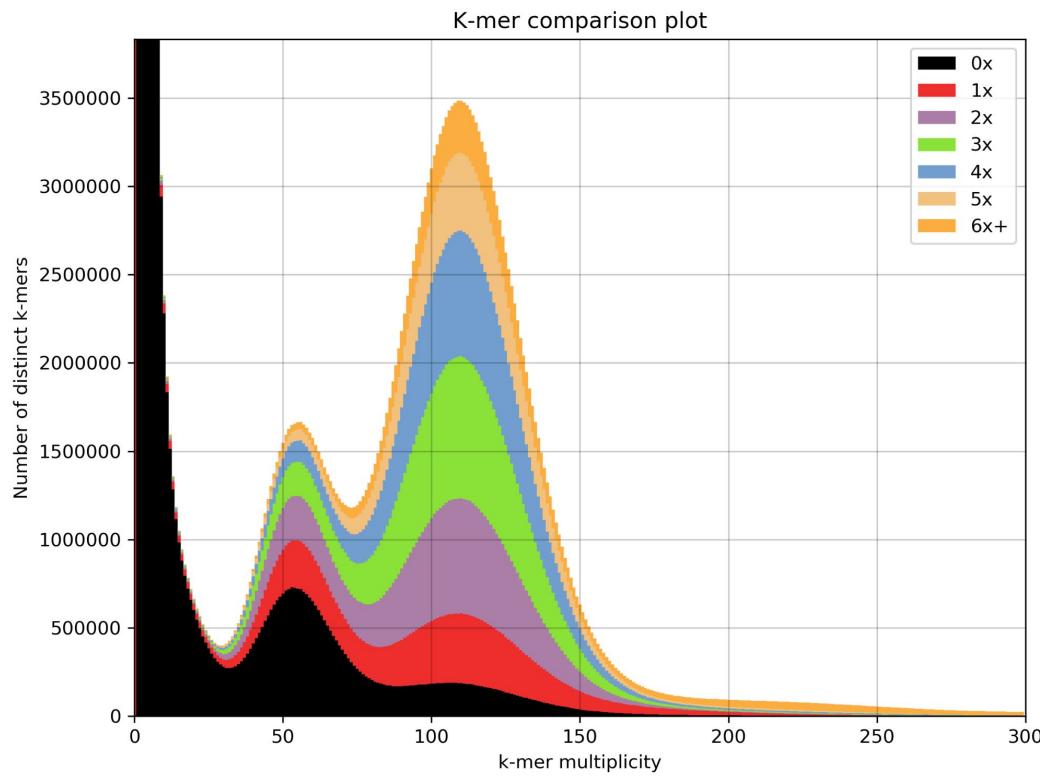
Merqury



k-mer completeness

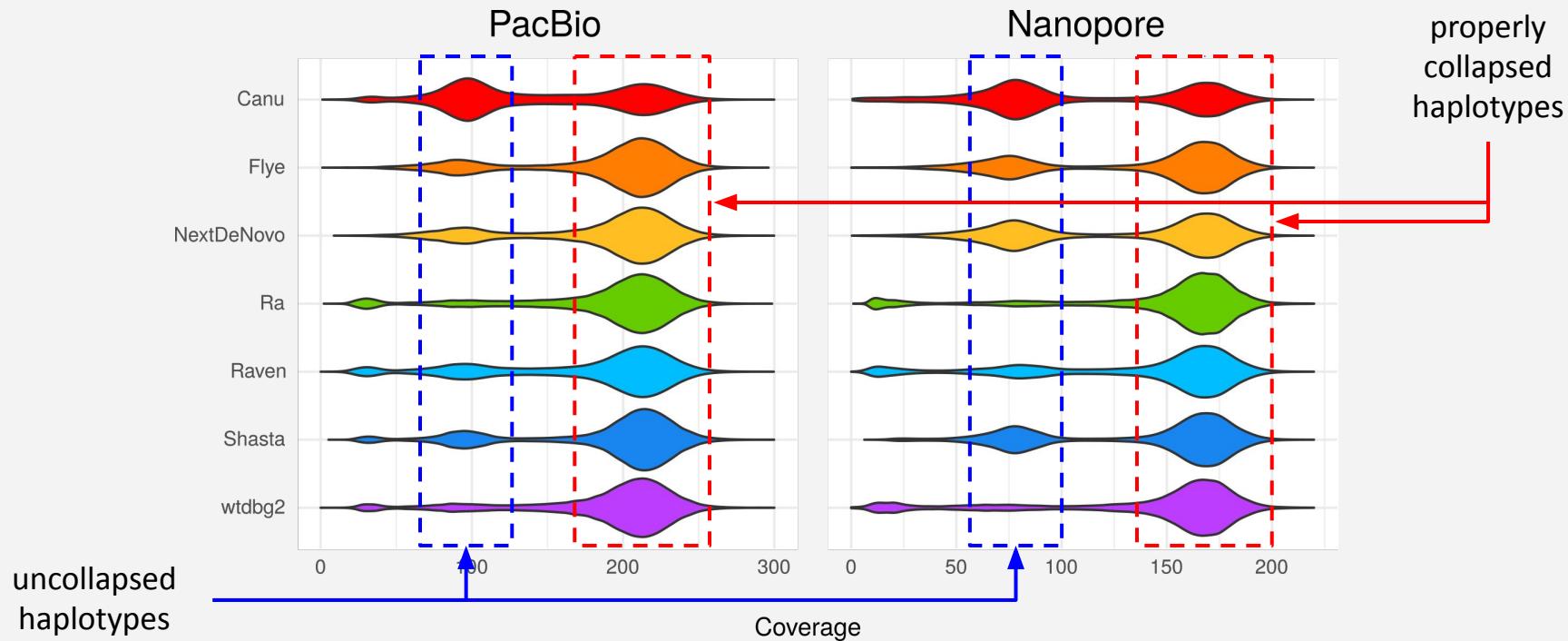
Gordionus montsenyensis

KAT comp



Coverage distribution

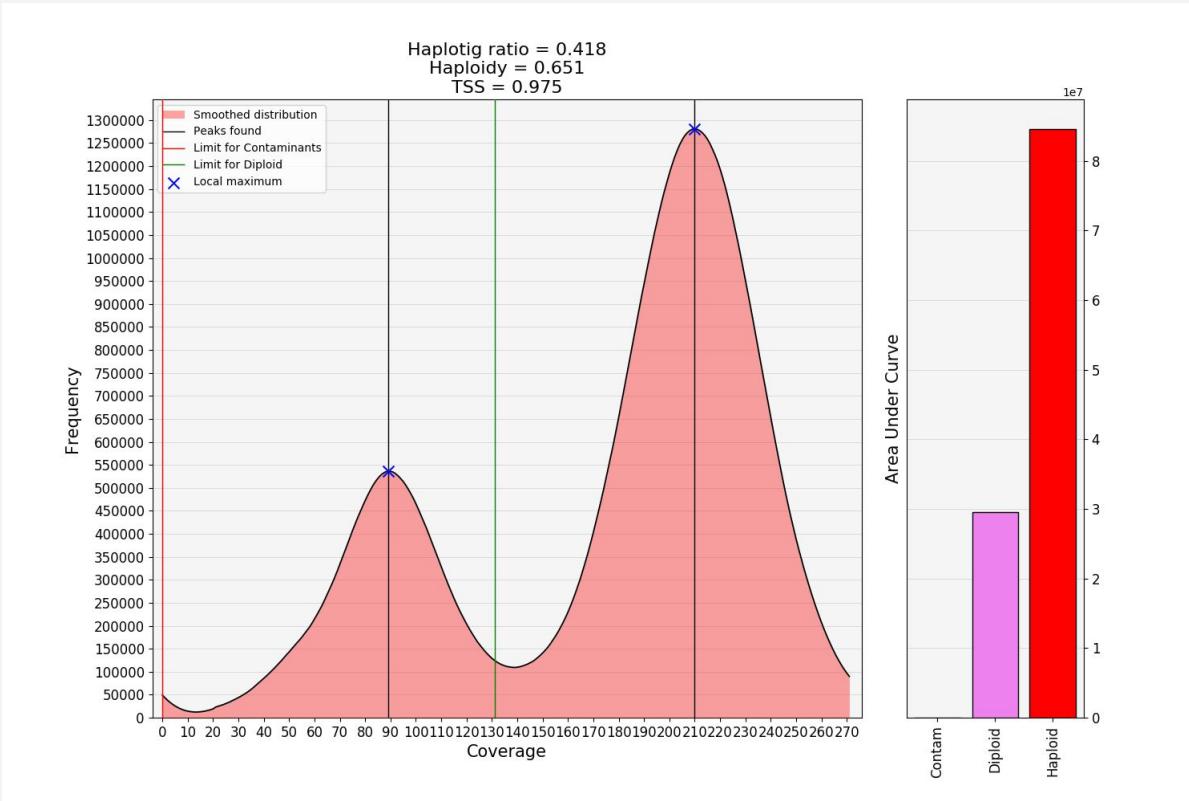
Adineta vaga



Coverage distribution

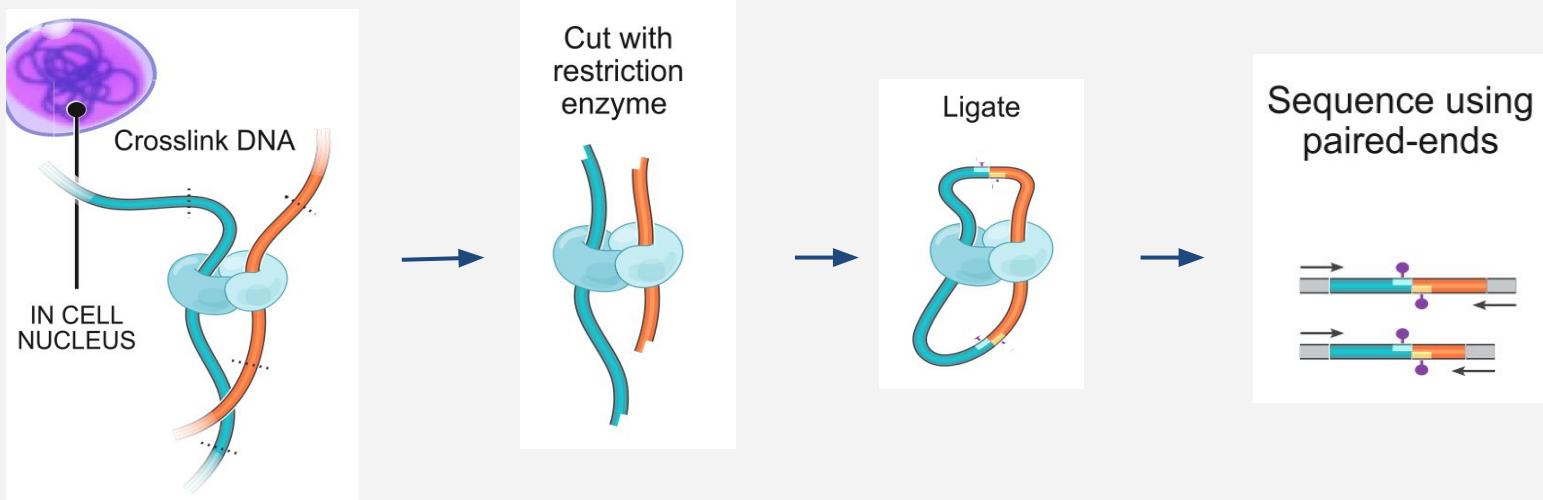
Adineta vaga

HapPy



Hi-C contact map

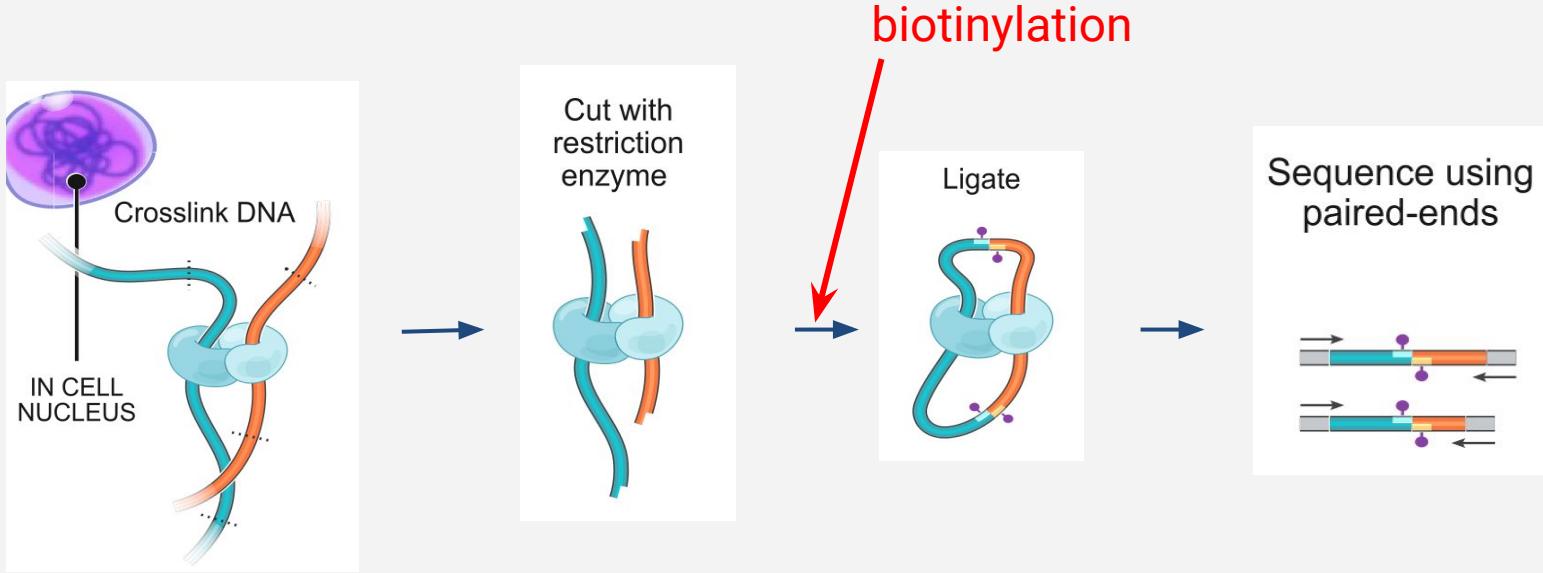
3C



A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Rao et al., 2014

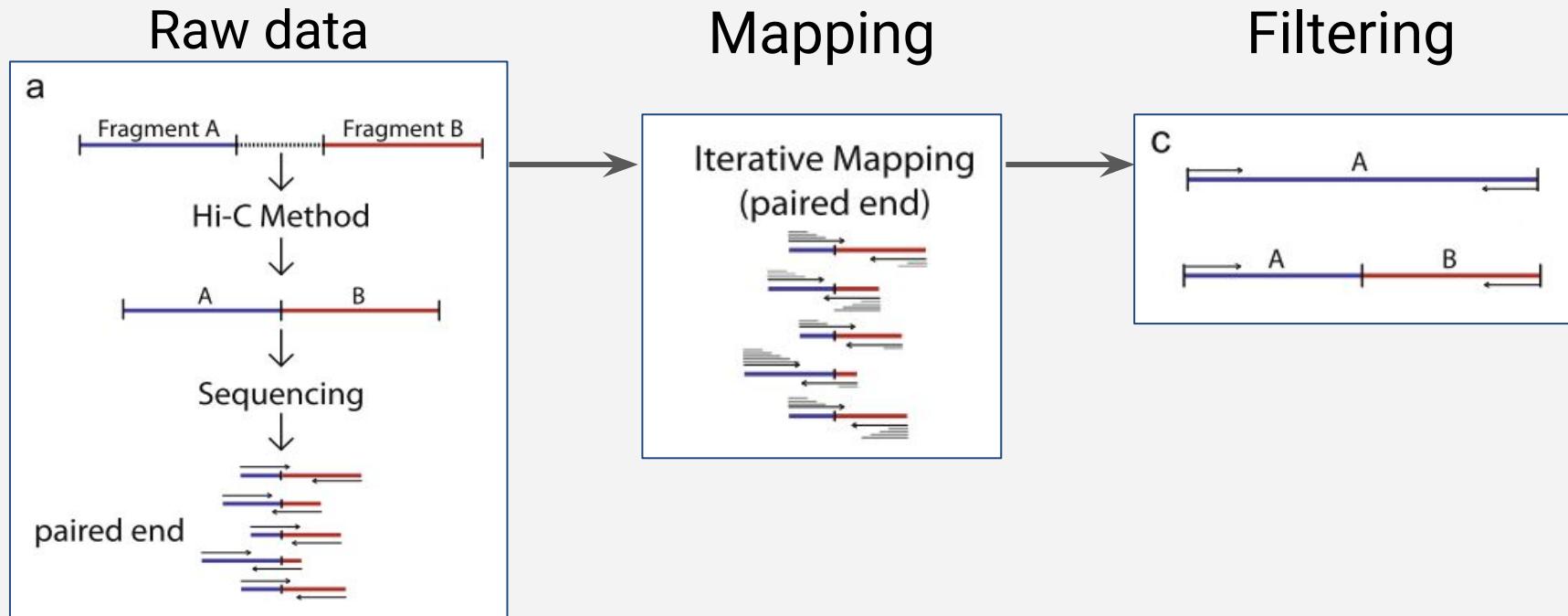
Hi-C contact map

Hi-C



A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Rao et al., 2014

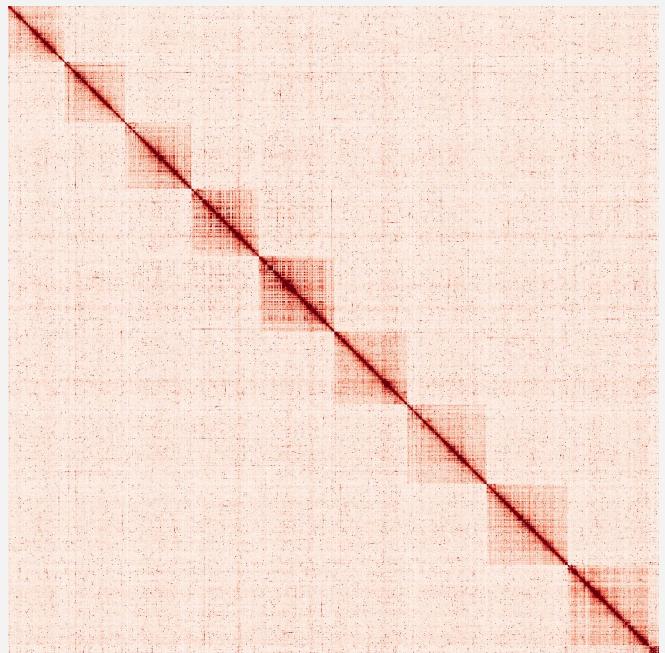
Hi-C contact map



Hi-C contact map

Flaccisagitta enflata

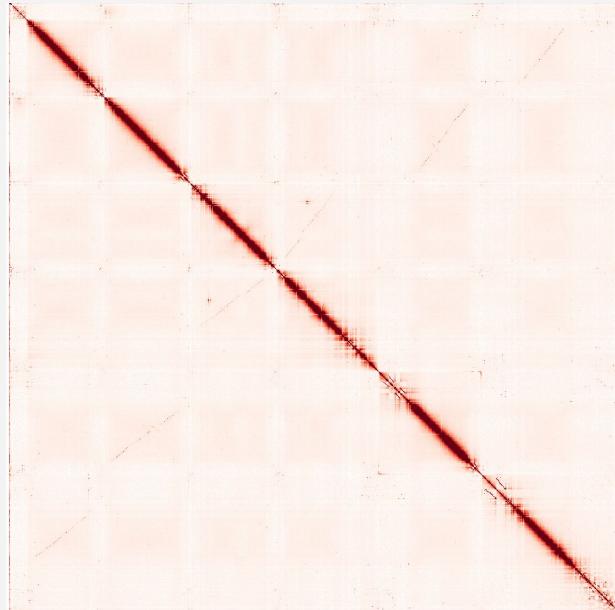
hicstuff



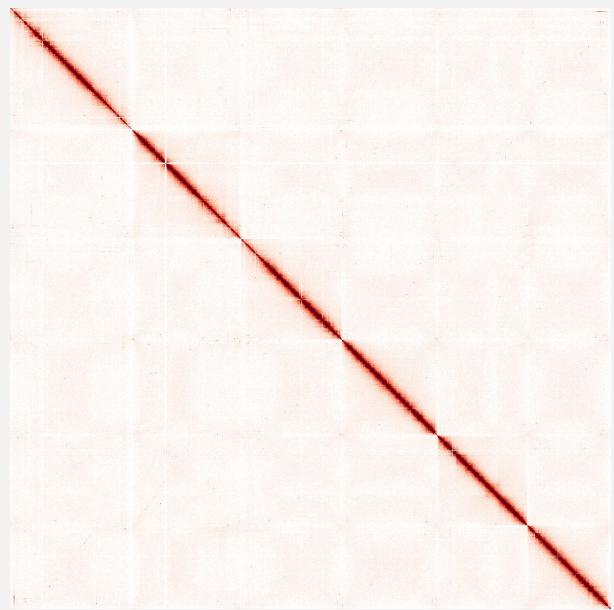
Hi-C contact map

Adineta vaga

hicstuff



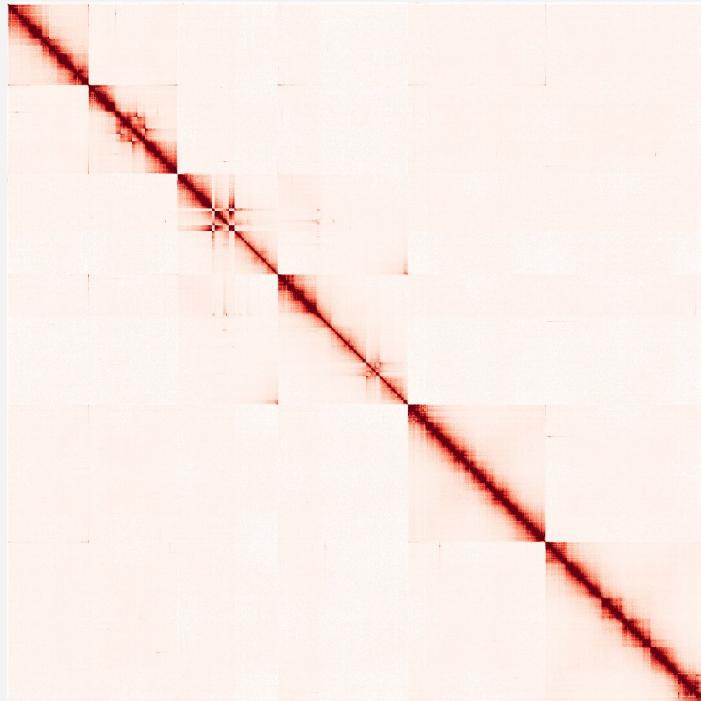
problematic assembly



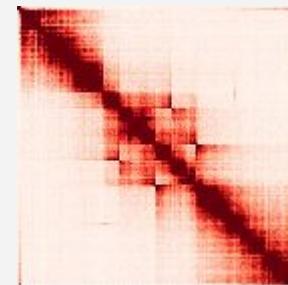
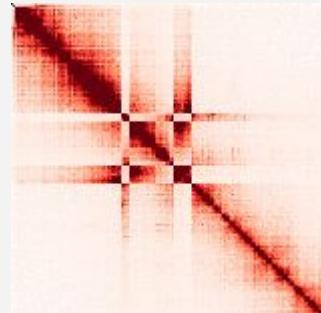
improved assembly

Hi-C contact map

Gordionus montsenyensis



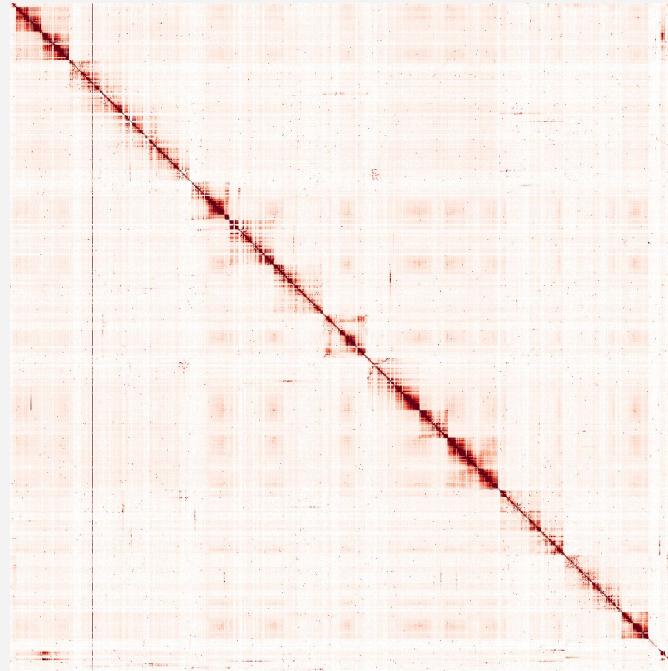
hicstuff



Hi-C contact map

Xenia sp.

hicstuff



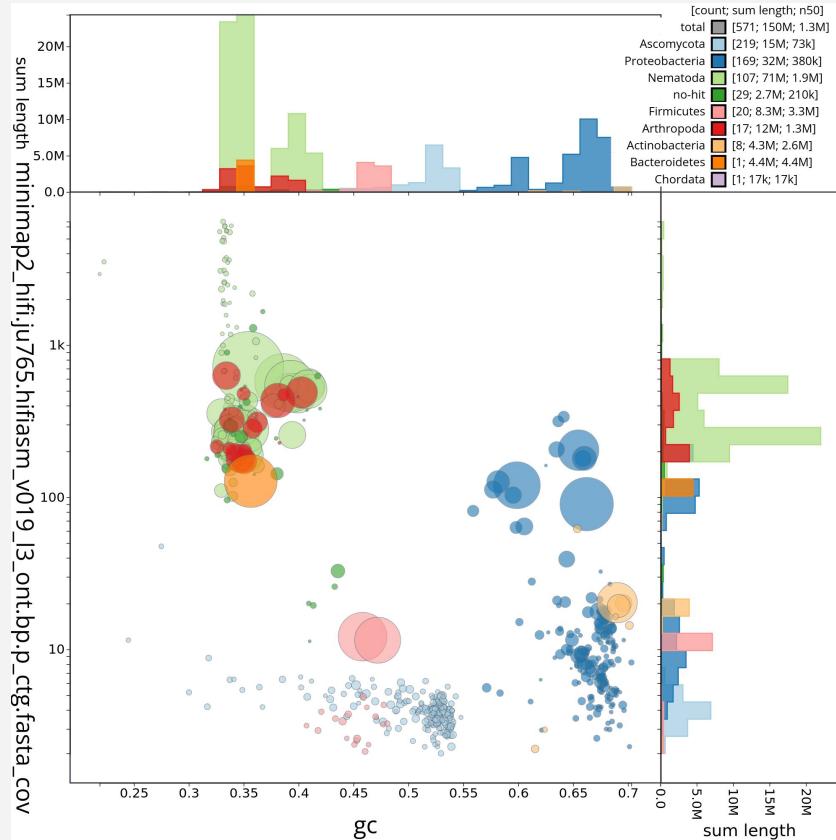
Super problematic

Contaminants

Propanagrolaimus JU765

Blobtools

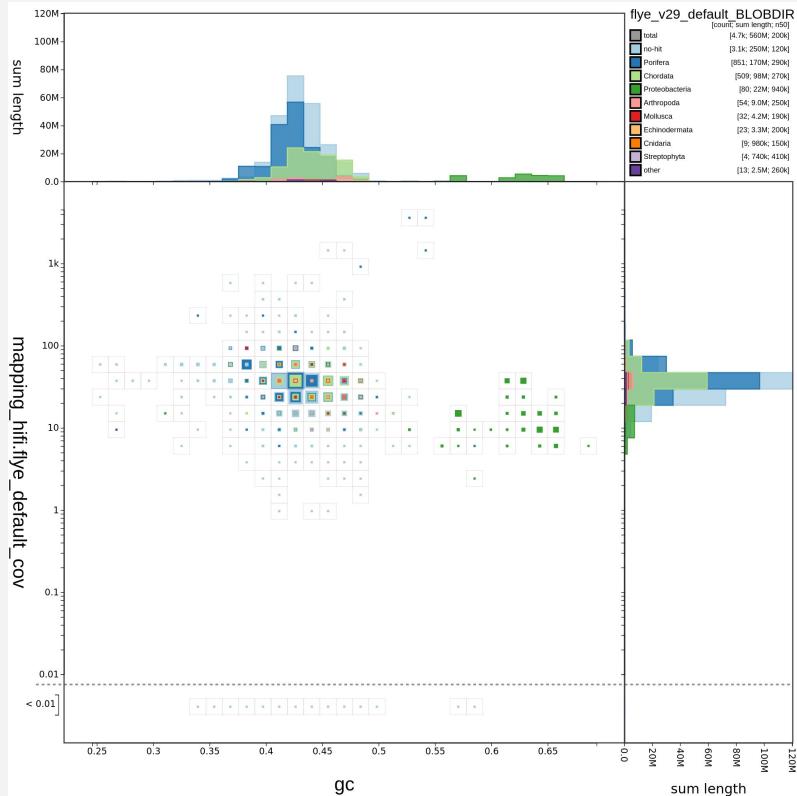
- Taxonomy based on BLAST/DIAMOND search
- Orthologs (BUSCO)
- Coverage depth



Contaminants

Spongilla lacustris

Blobtools



Tools to play with

- ▶ *k*-mer analysis: GenomeScope*, KAT, Merqury*, smudgeplot
- ▶ Reads quality control: NanoPlot*, fastqc
- ▶ Ortholog completeness: BUSCO*
- ▶ N50: gfastats*, assembly-stats, QUAST
- ▶ Assembly graph: Bandage*
- ▶ Coverage distribution: HapPy