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

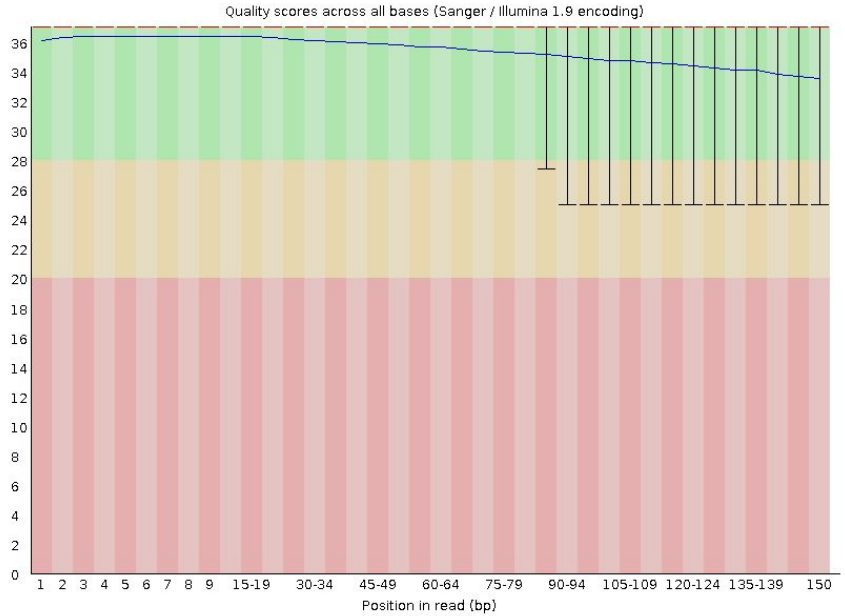
Sequencing data analysis

Illumina fastqc

Summary

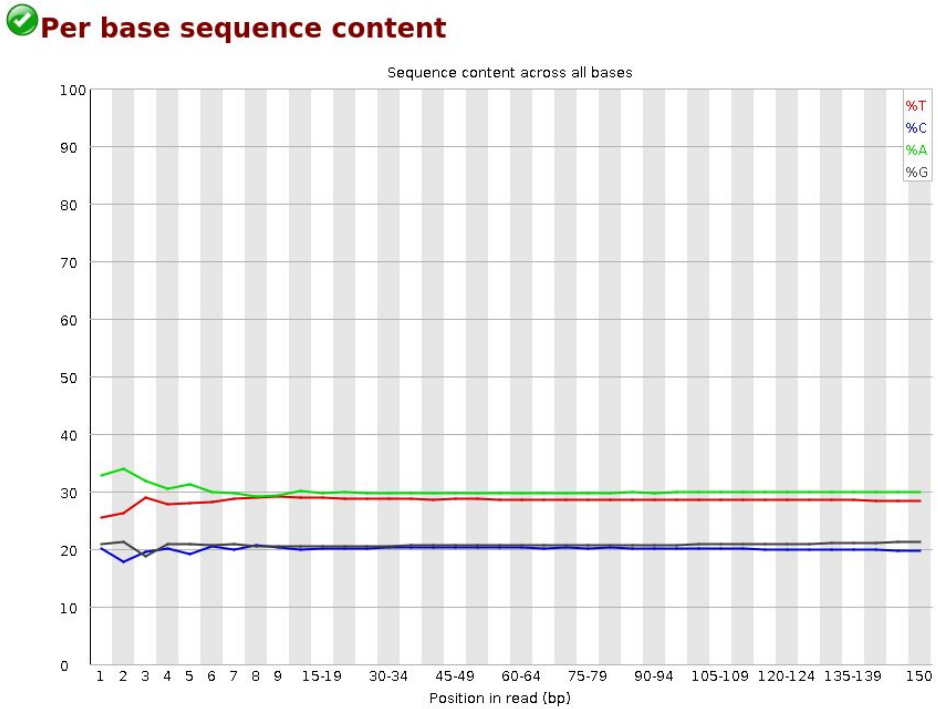
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Per base sequence quality



Sequencing data analysis

Illumina fastqc



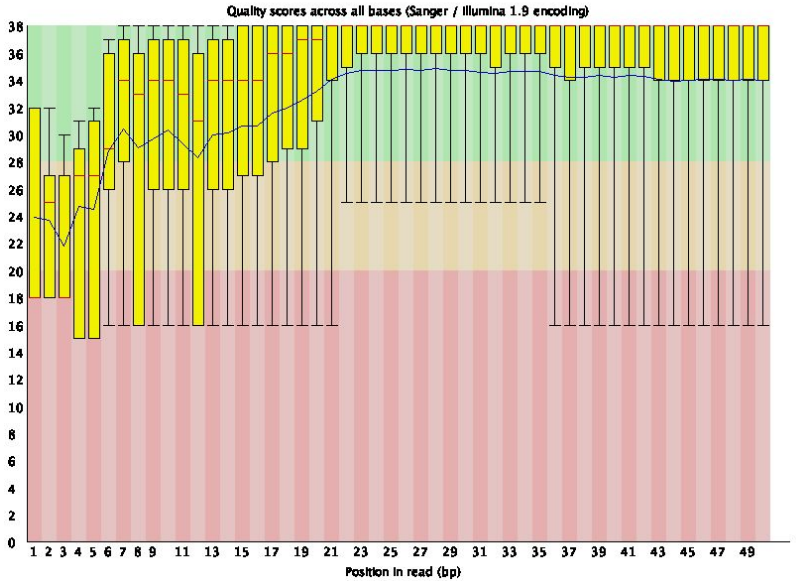
Sequencing data analysis

Illumina fastqc

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

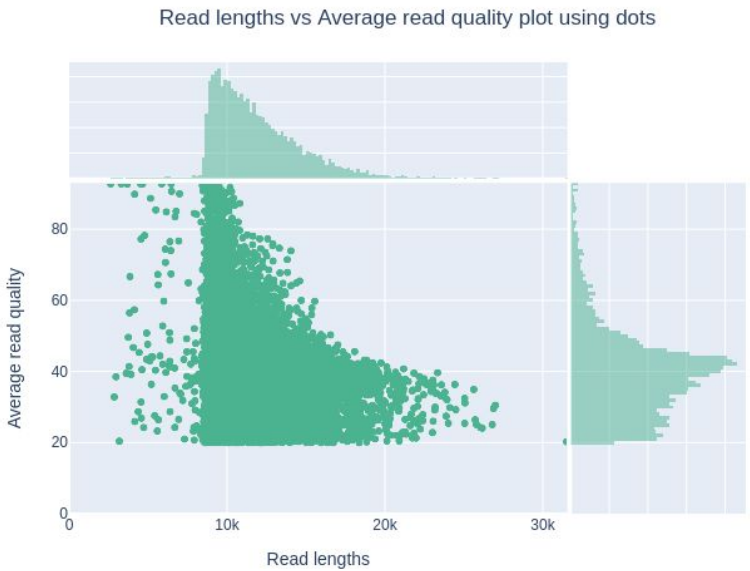
Per base sequence quality



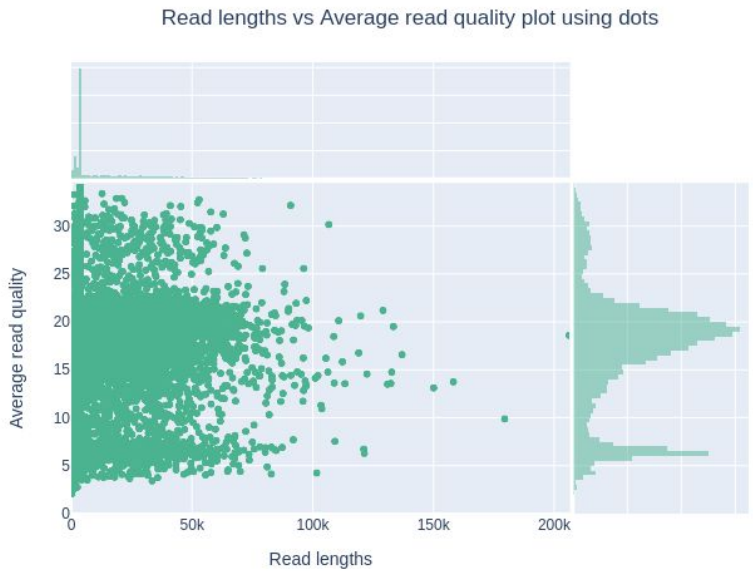
Sequencing data analysis

NanoPlot

PacBio HiFi



Nanopore

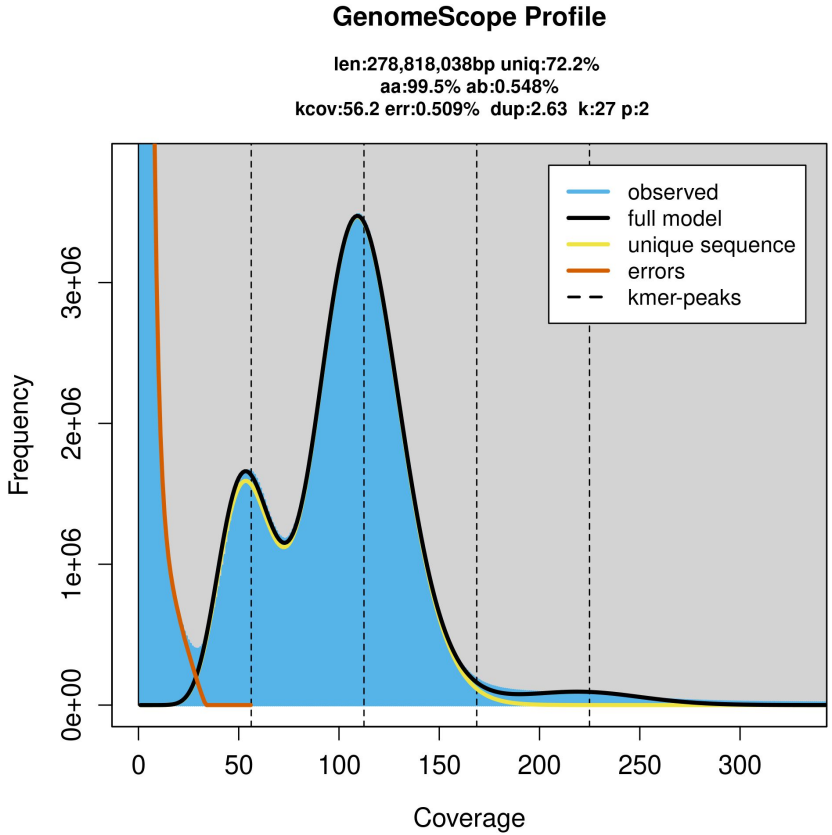


Sequencing data analysis

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

Sequencing data analysis

Gordionus montsenyensis

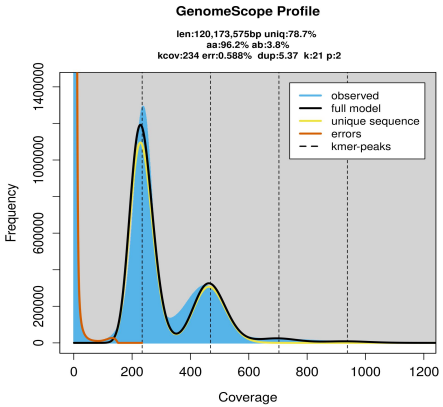


The genome sequence of the Montseny horsehair worm, *Gordionus montsenyensis* sp. nov., a key resource to investigate Ecdysozoa evolution. Eleftheriadi et al. bioRxiv, 2023.

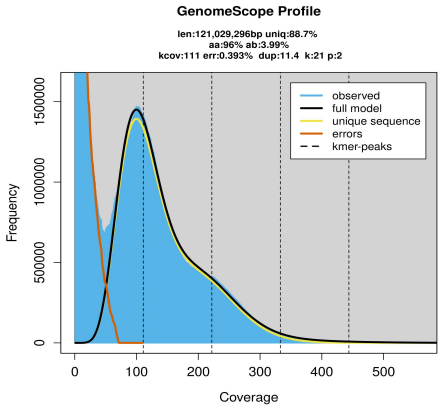
Sequencing data analysis

Plectus sambesii

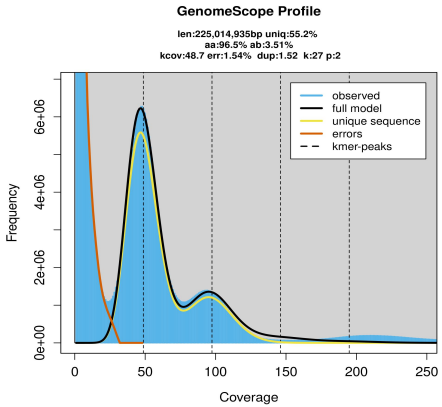
Illumina



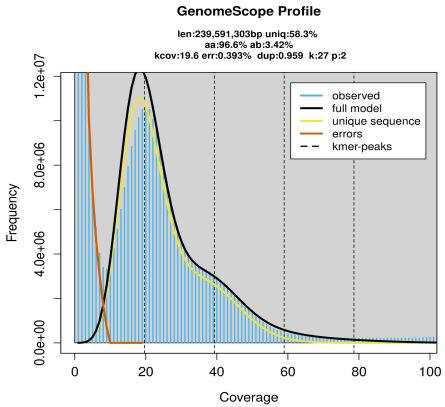
PacBio HiFi



Nanopore

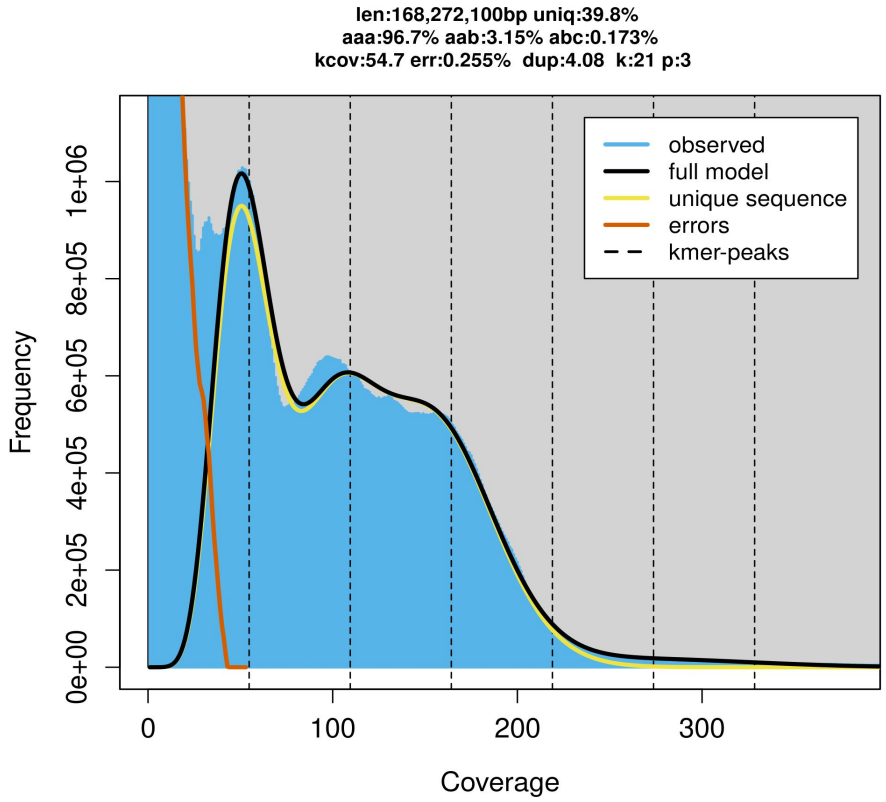


Nanopore Q20



Sequencing data analysis

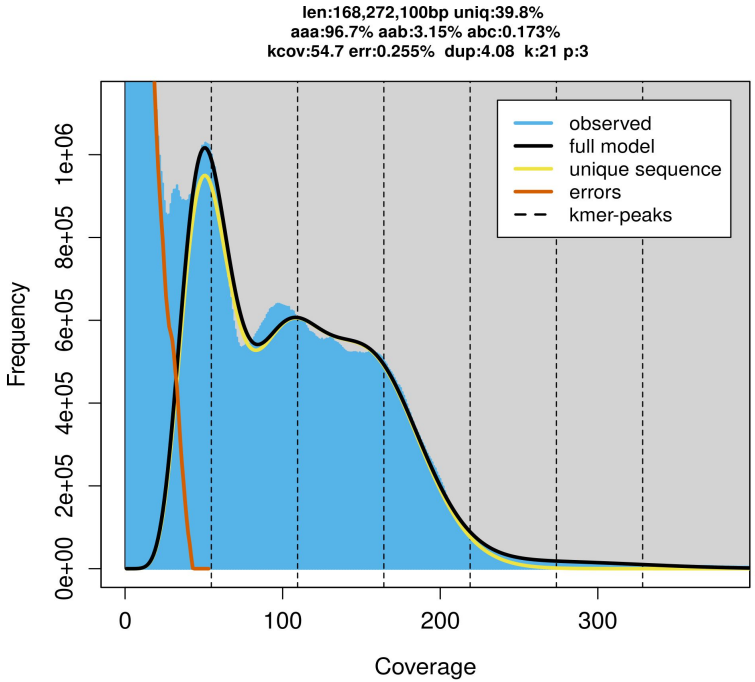
Panagrolaimus PS1159



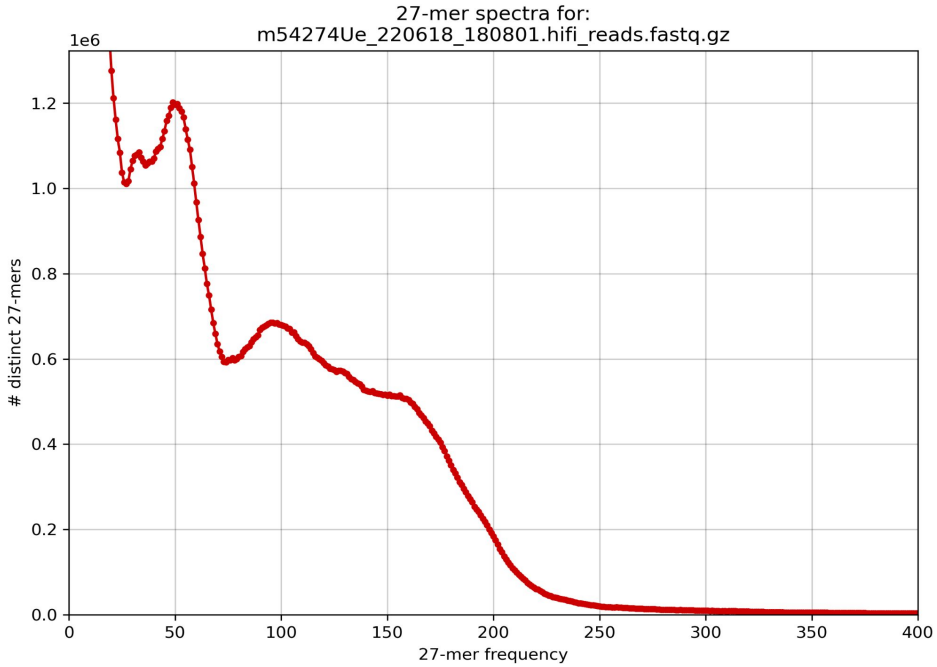
Sequencing data analysis

Panagrolaimus PS1159

GenomeScope



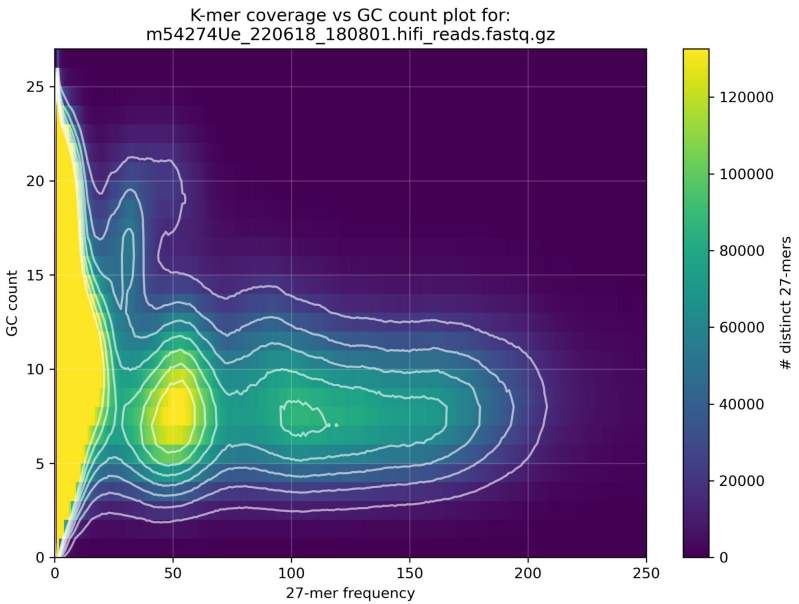
KAT hist



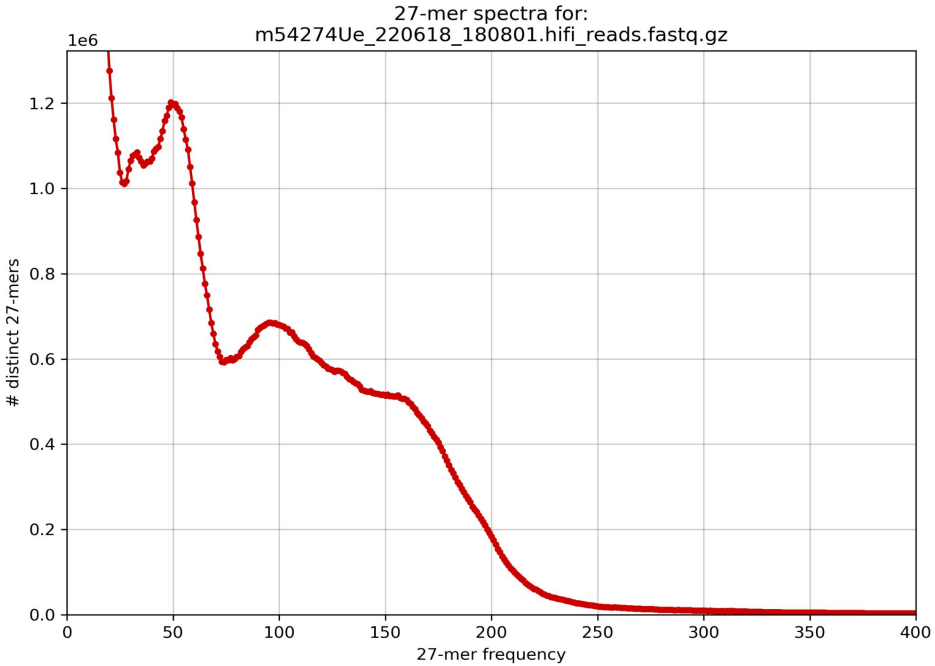
Sequencing data analysis

Panagrolaimus PS1159

KAT gcp



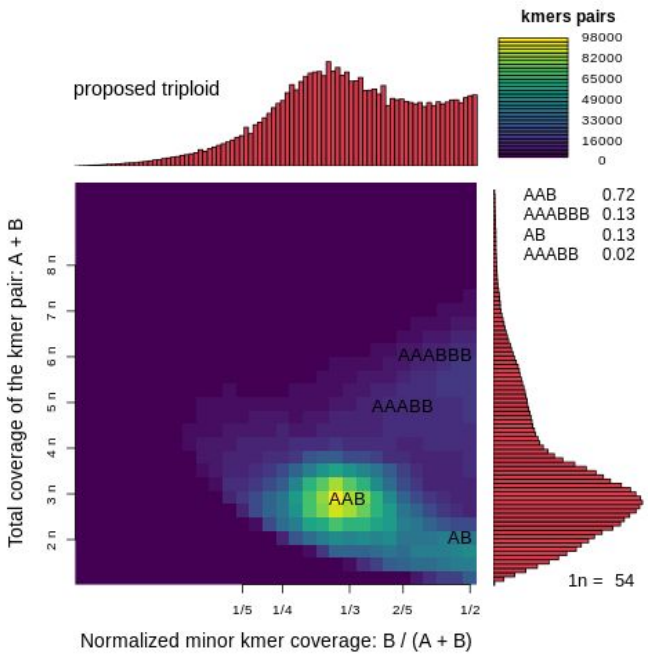
KAT hist



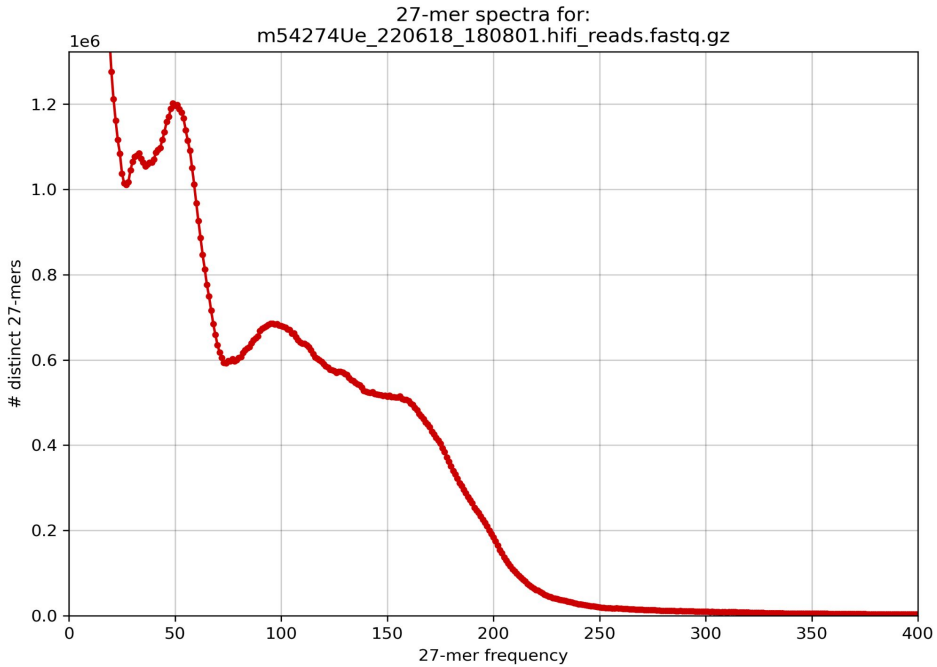
Sequencing data analysis

Panagrolaimus PS1159

Smudgeplot

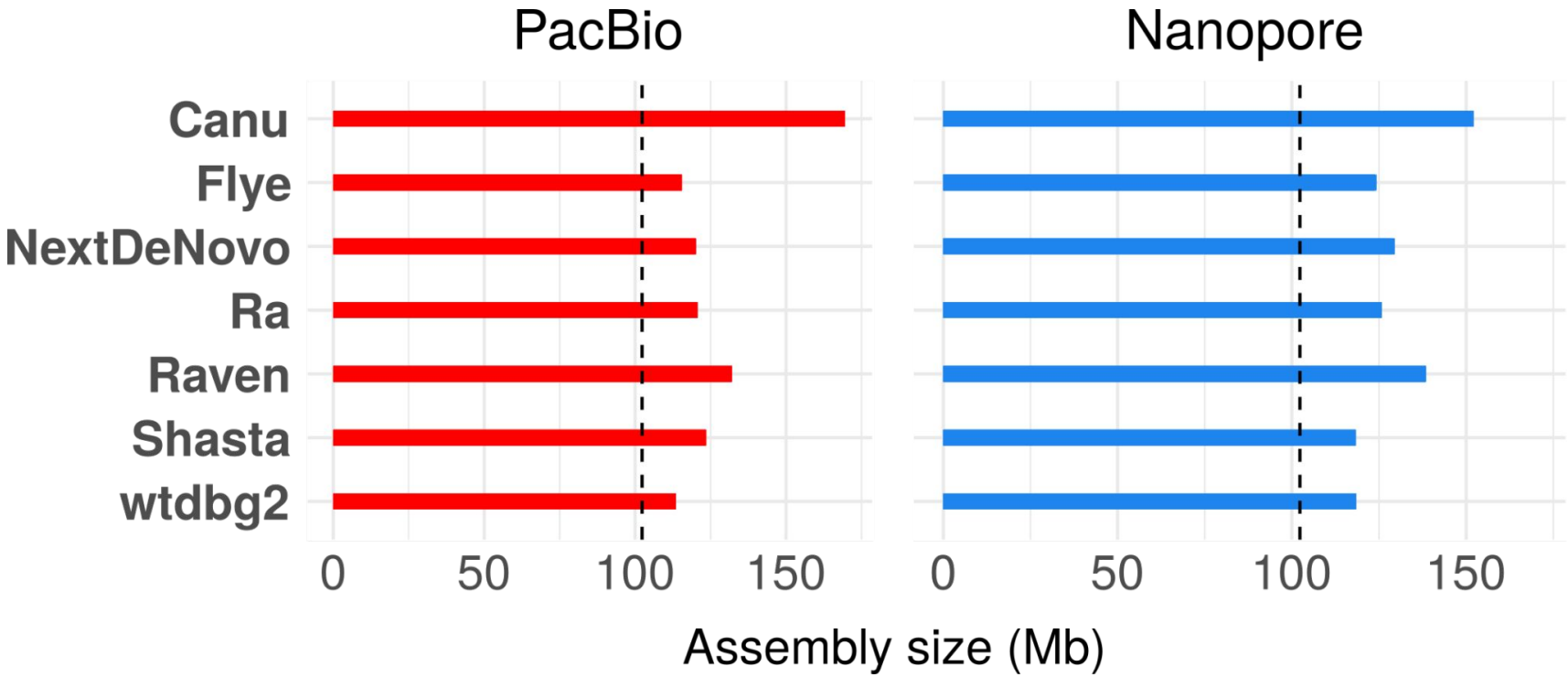


KAT hist



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**

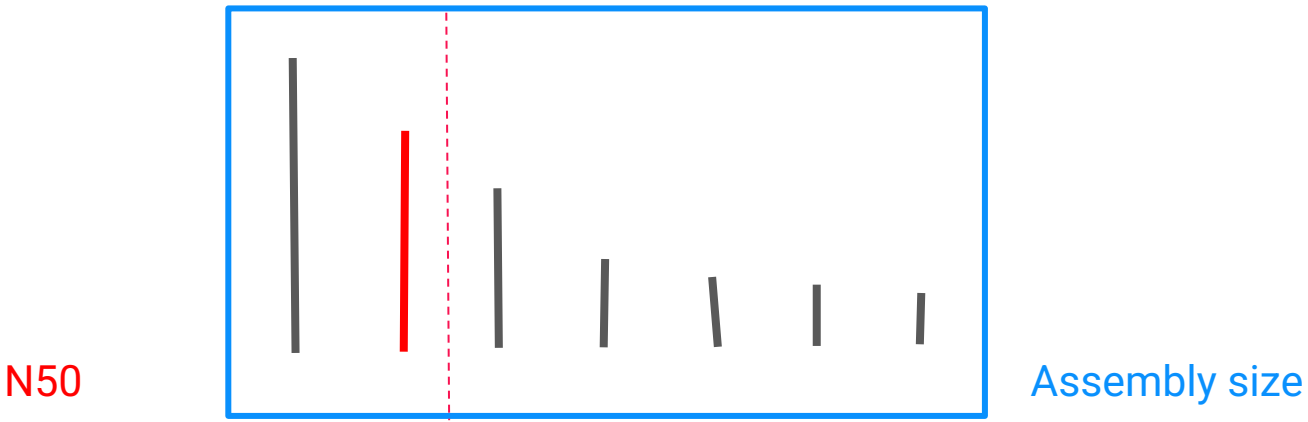
Contiguity

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**

Assembly size

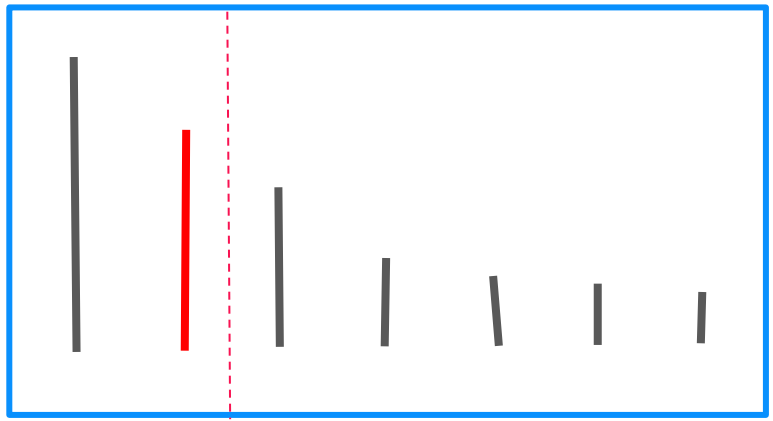
Contiguity

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**

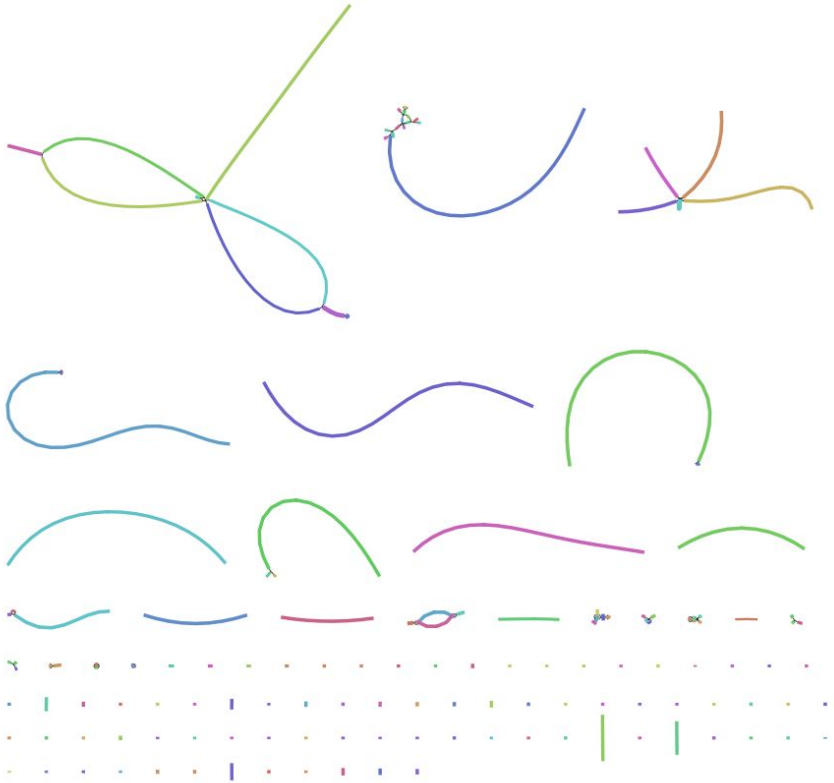


N50
L50 = 2

Assembly size

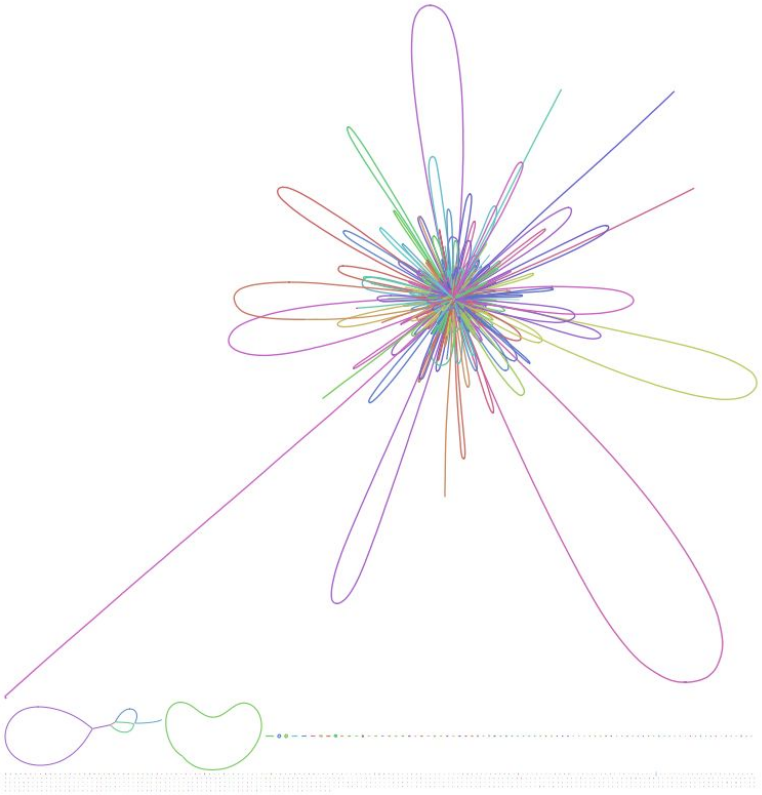
Assembly graph

Bandage

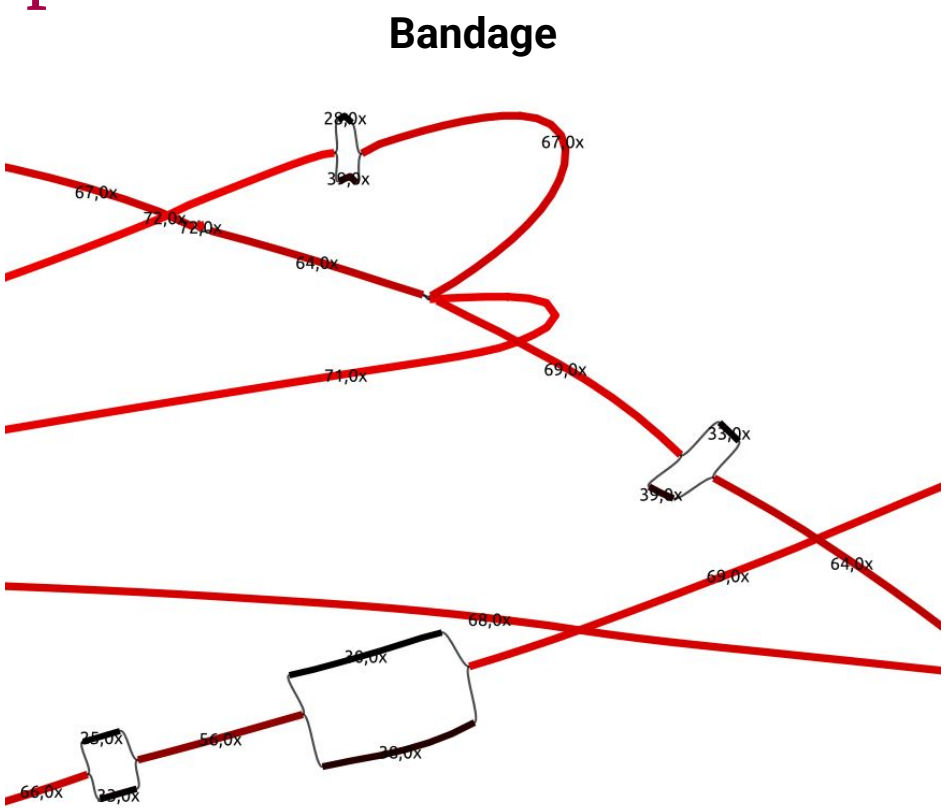


Assembly graph

Bandage



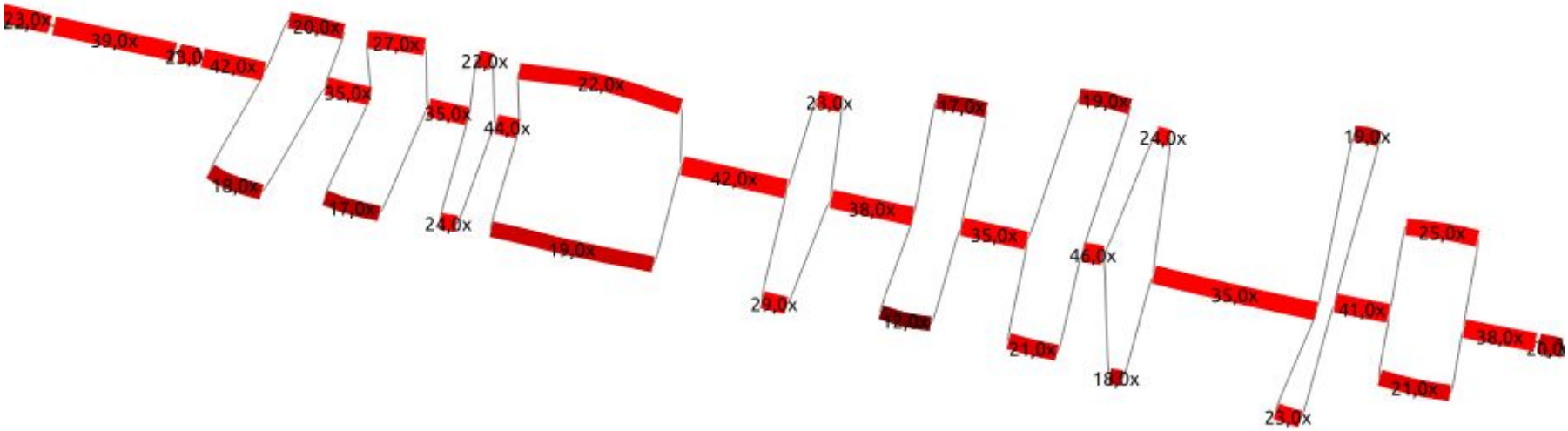
Assembly graph



bubbles

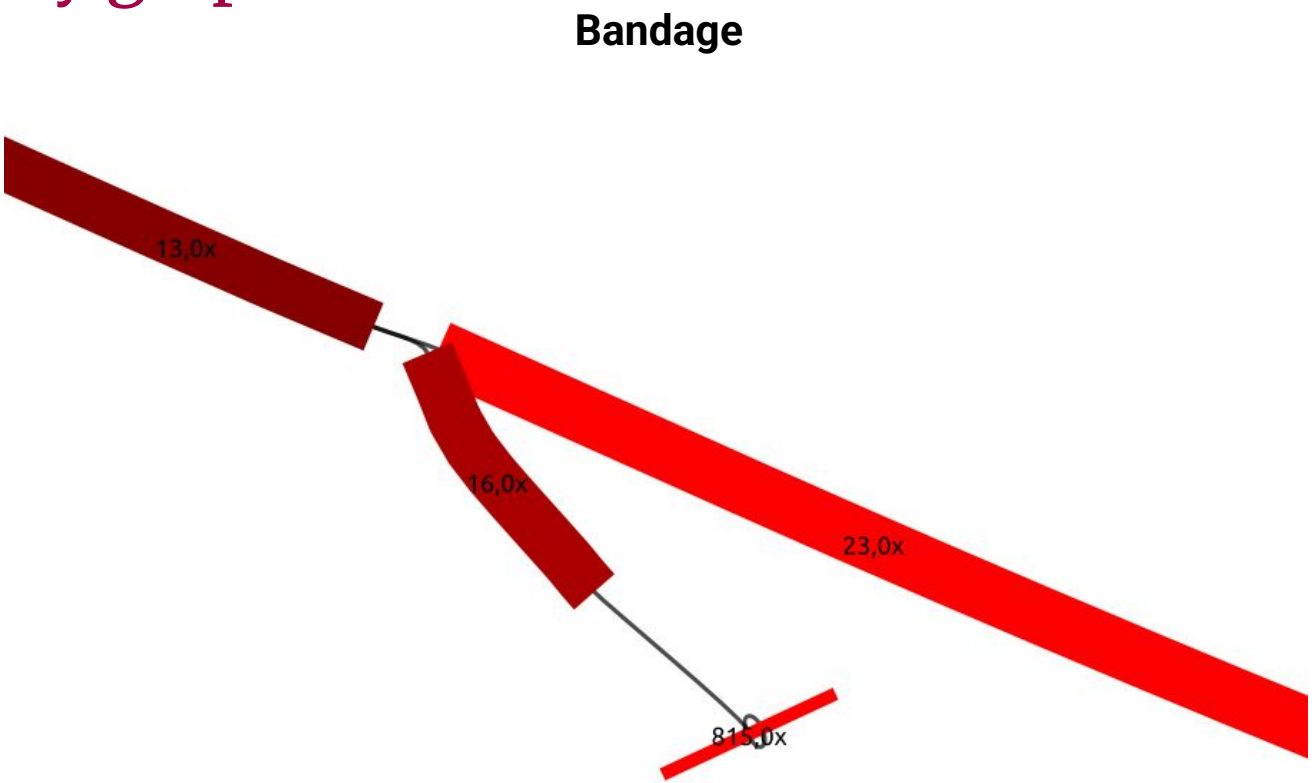
Assembly graph

Bandage



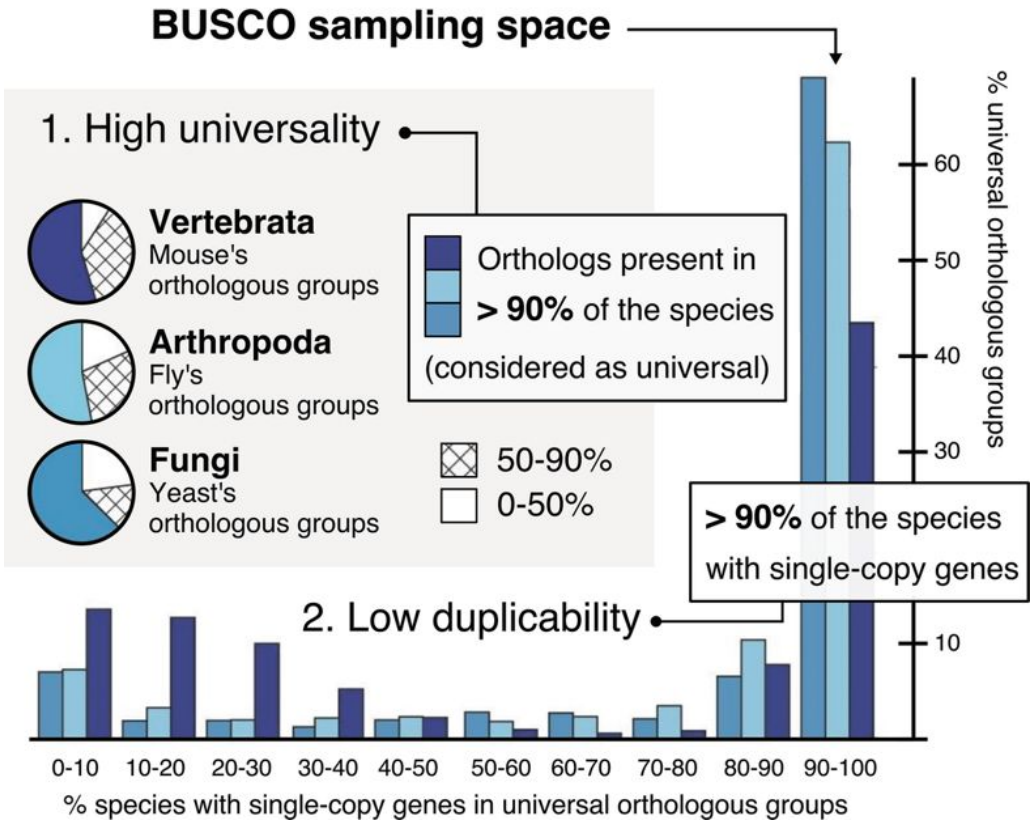
bubbles

Assembly graph



repetitions

Ortholog completeness



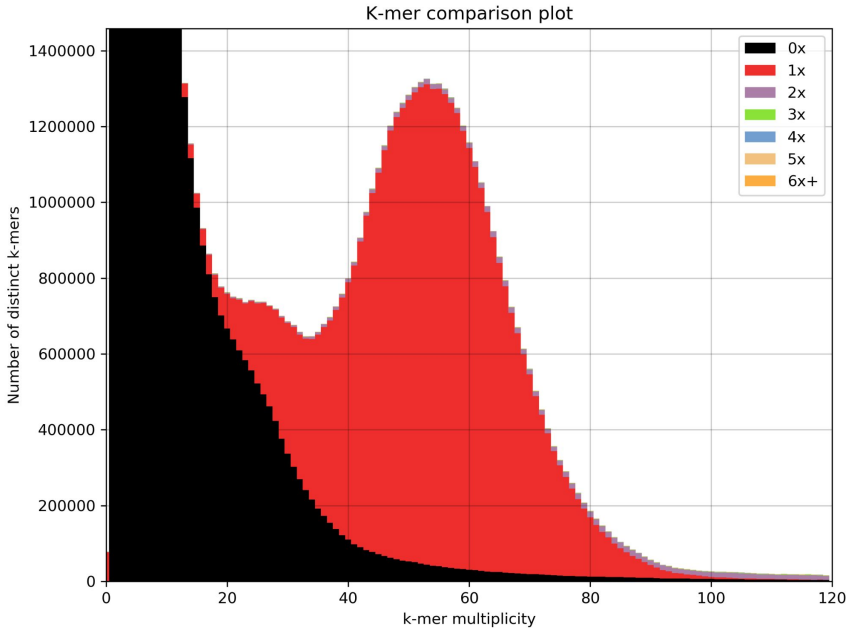
Ortholog completeness

Adineta vaga

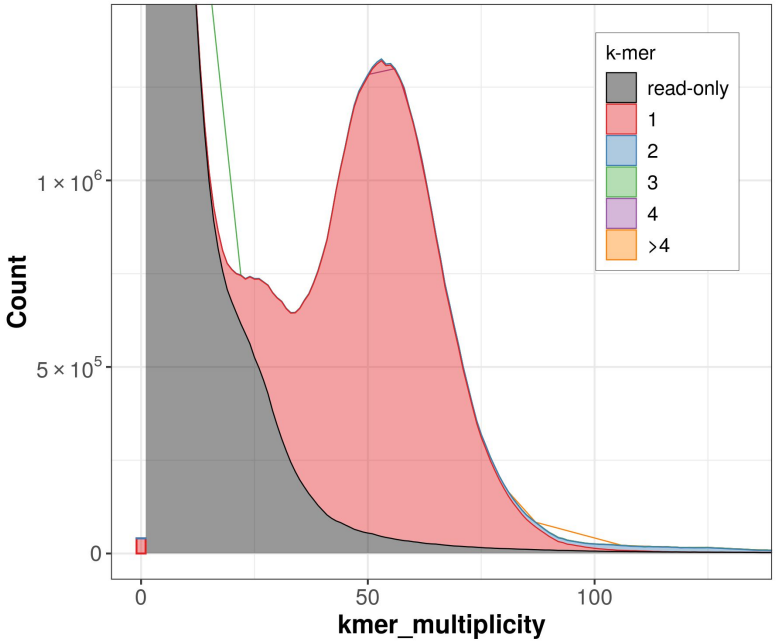
Reads	Assembly	Complete features	Complete single-copy	Complete duplicated	Fragmented	Missing
Illumina	BWISE	82.6%	16.4%	66.2%	3.2%	14.2%
PacBio	NextDenovo	84.1%	72.0%	12.1%	2.8%	13.1%
Nanopore	NextDenovo	44.7%	39.6%	5.1%	21.1%	34.2%
Nanopore + Illumina	NextDenovo	86.2%	72.0%	14.2%	1.9%	11.9%
HiFi	hifiasm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

k-mer completeness

KAT comp

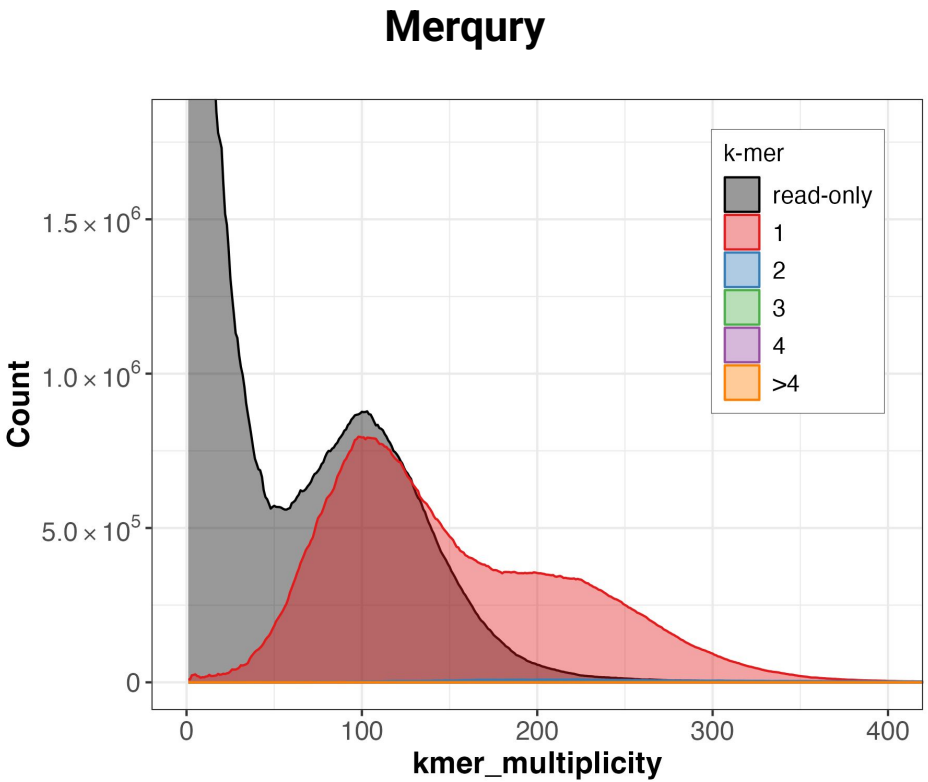


Merqury



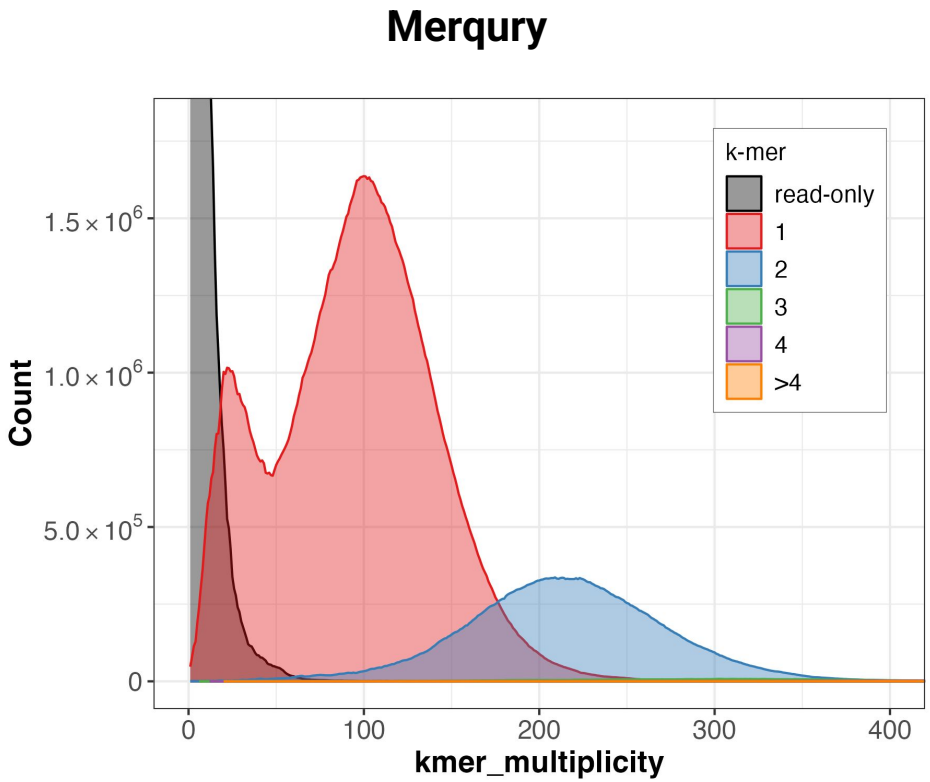
k-mer completeness

Plectus sambesii



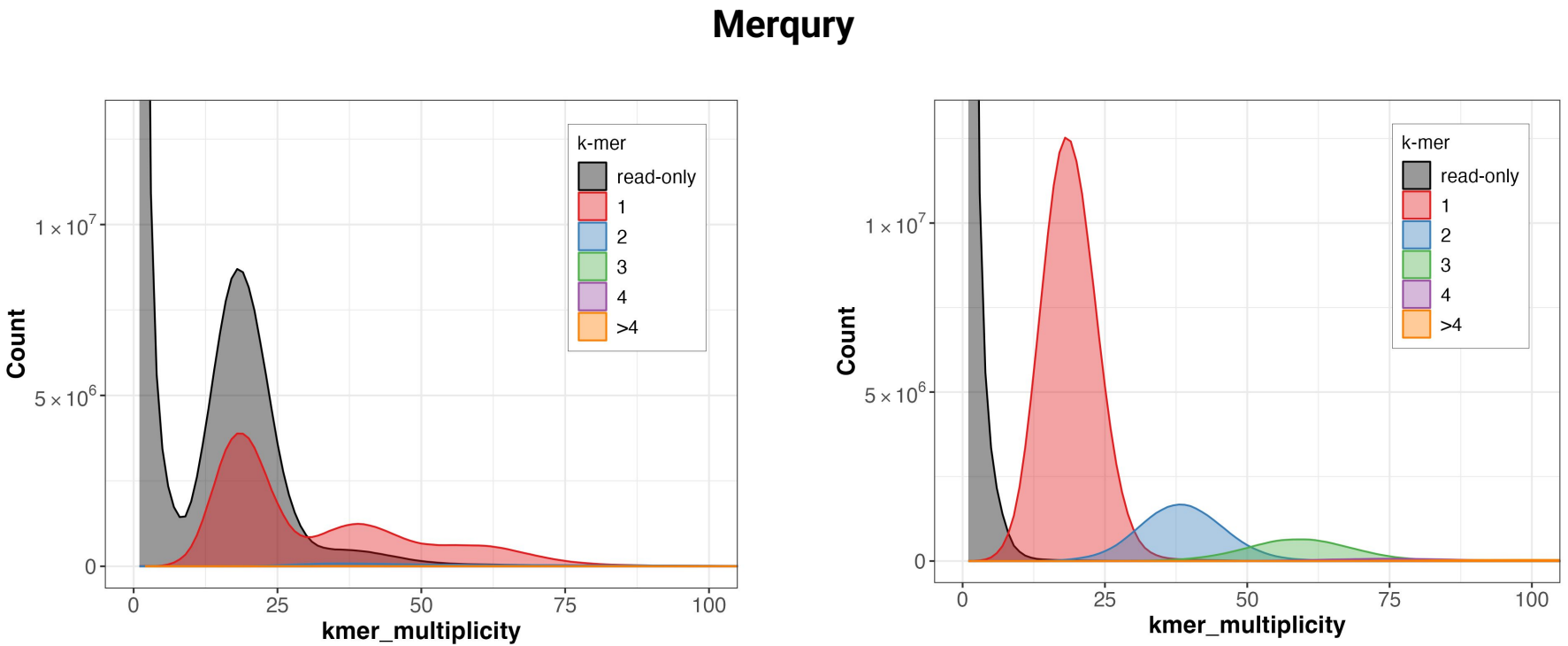
k-mer completeness

Plectus sambesii



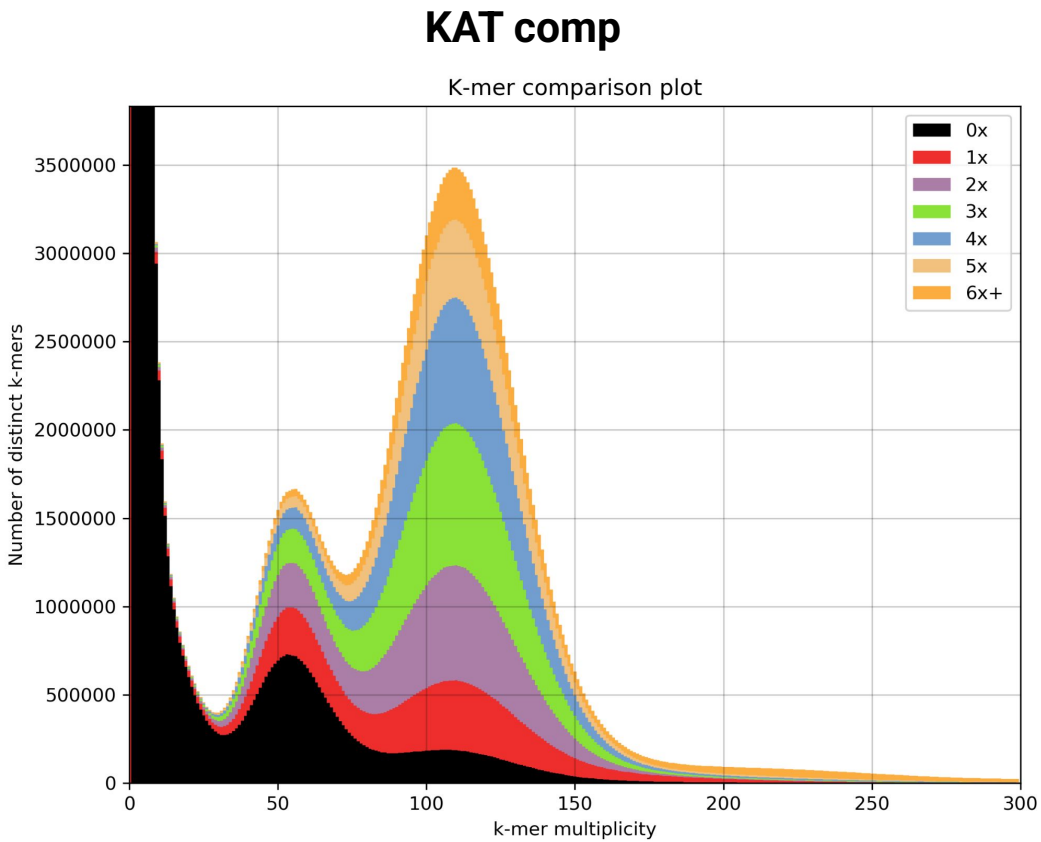
k-mer completeness

Panagrolaimus PS1579



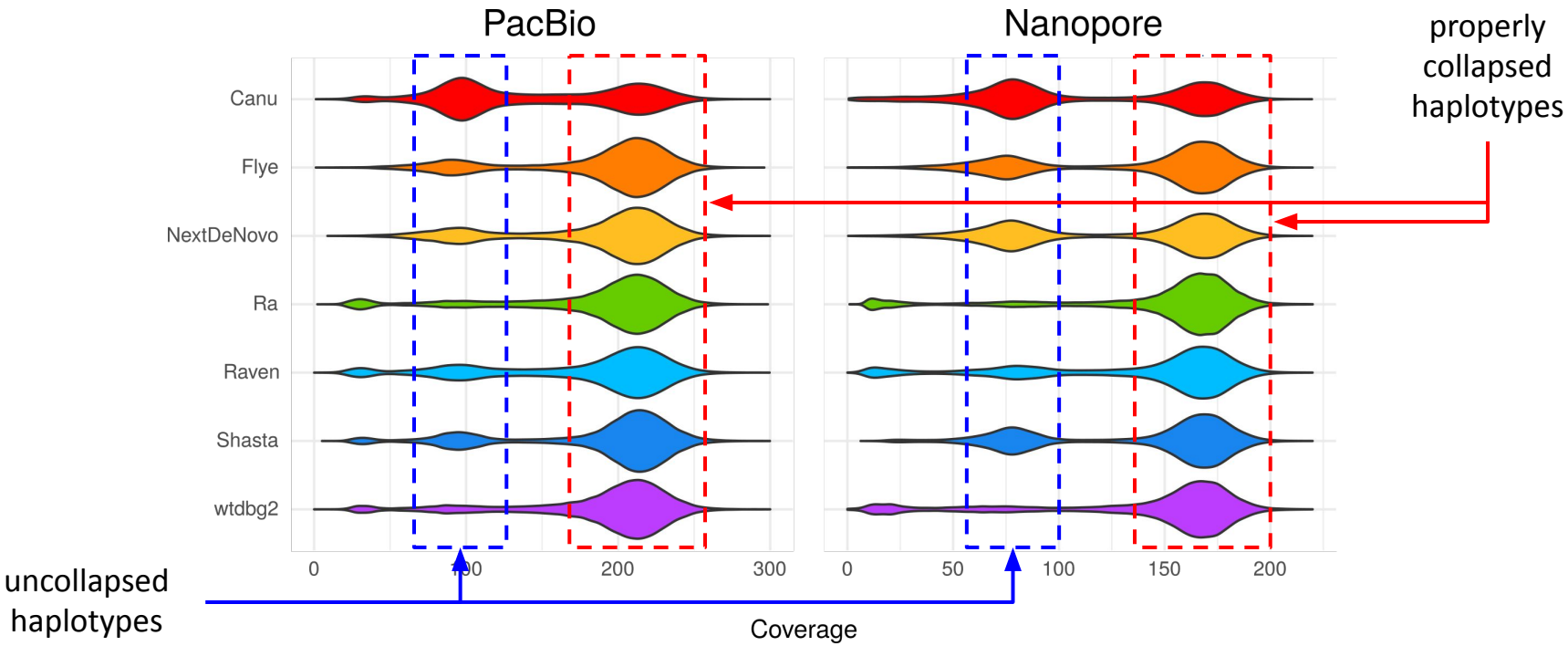
k-mer completeness

Gordionus montsenyensis



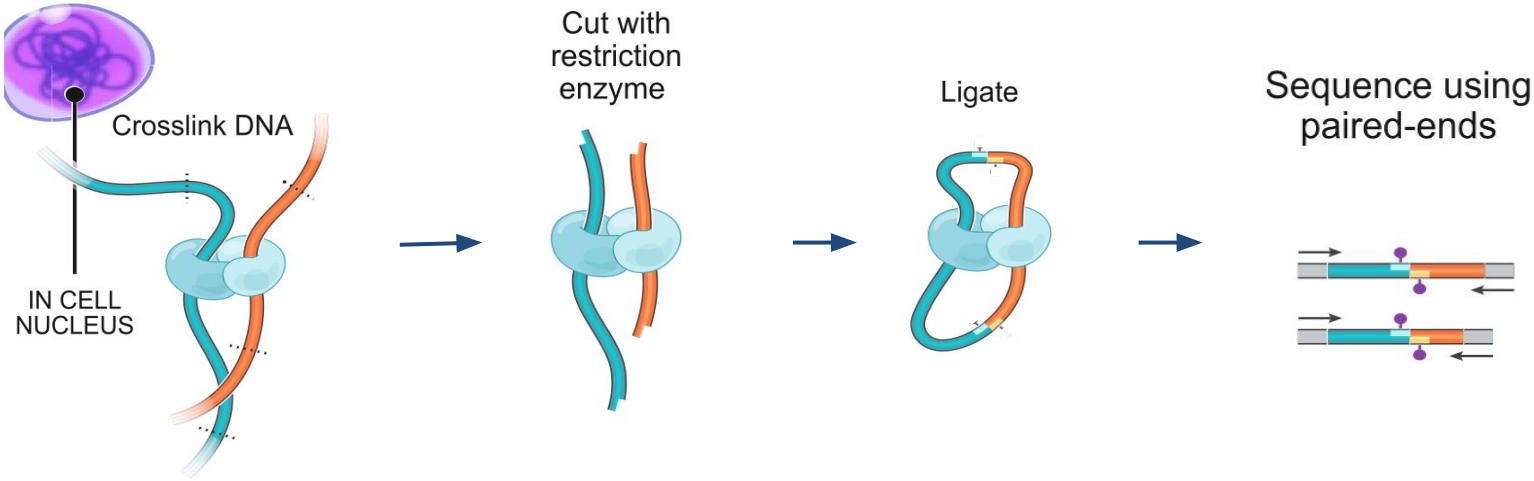
Coverage distribution

Adineta vaga



Hi-C contact map

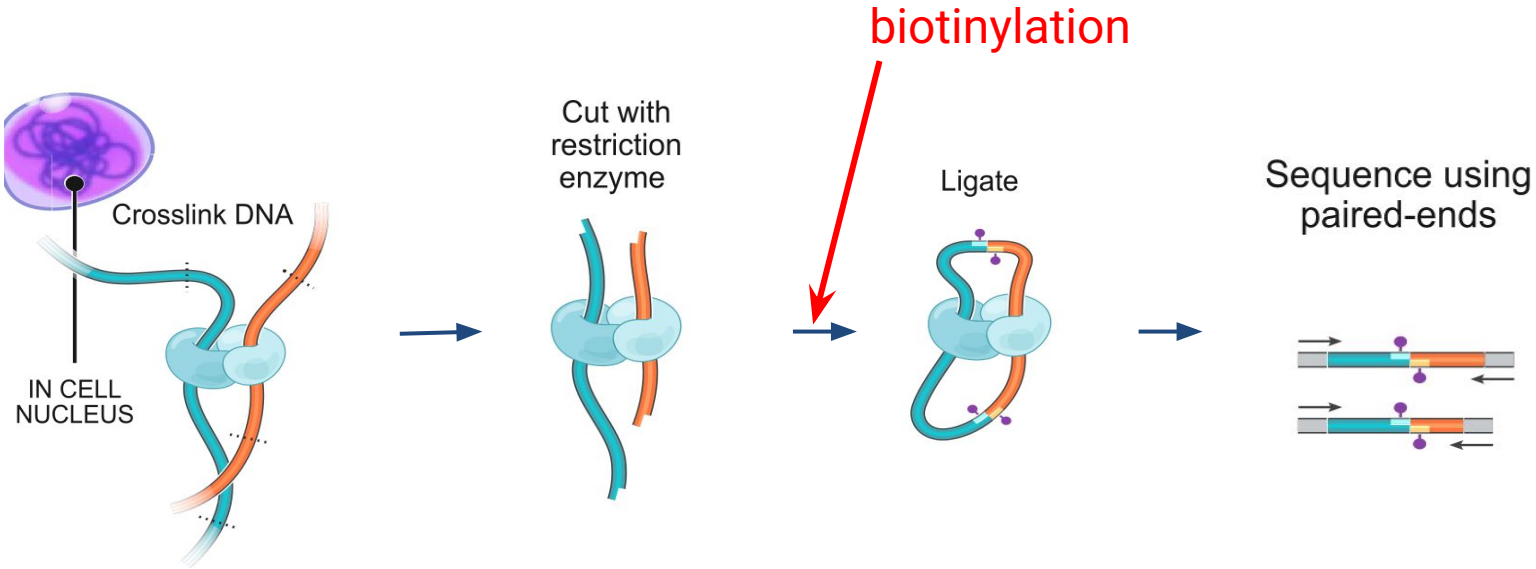
3C



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

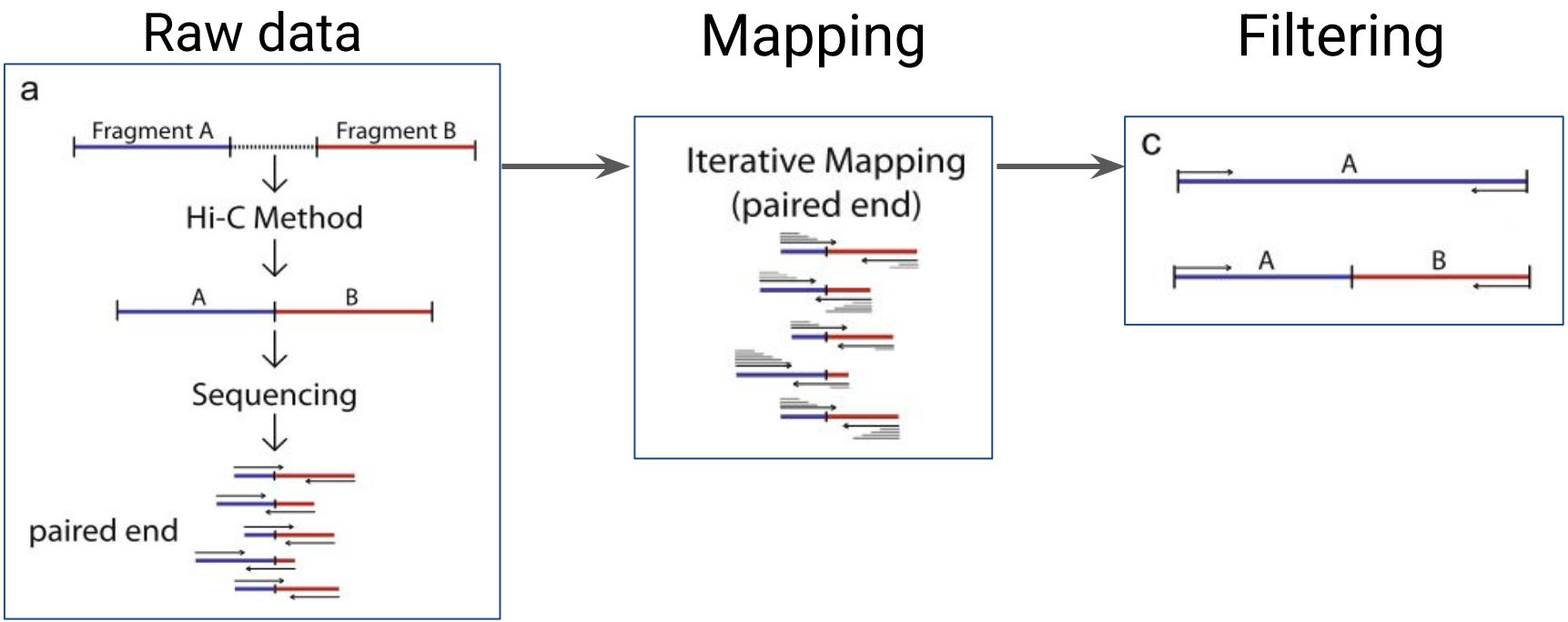
Hi-C contact map

Hi-C



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

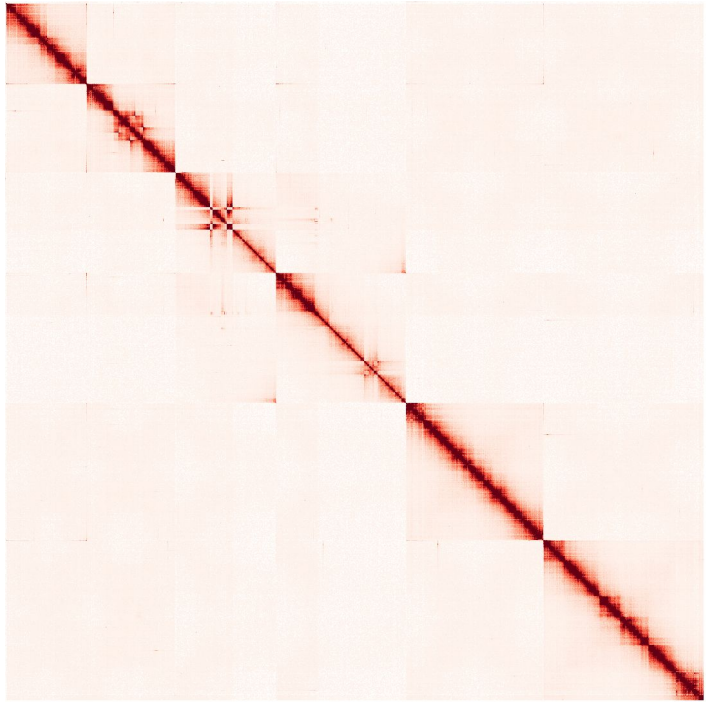
Hi-C contact map



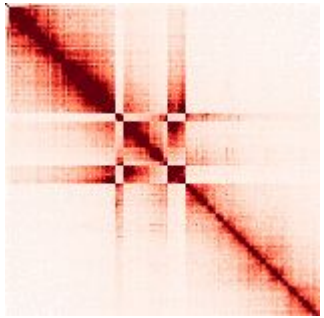
The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Lajoie et *al.*, 2015

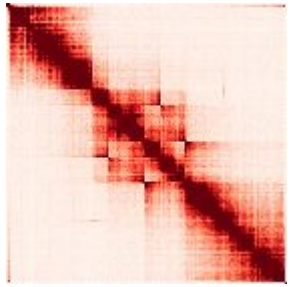
Hi-C contact map

Gordionus montsenyensis



hicstuff

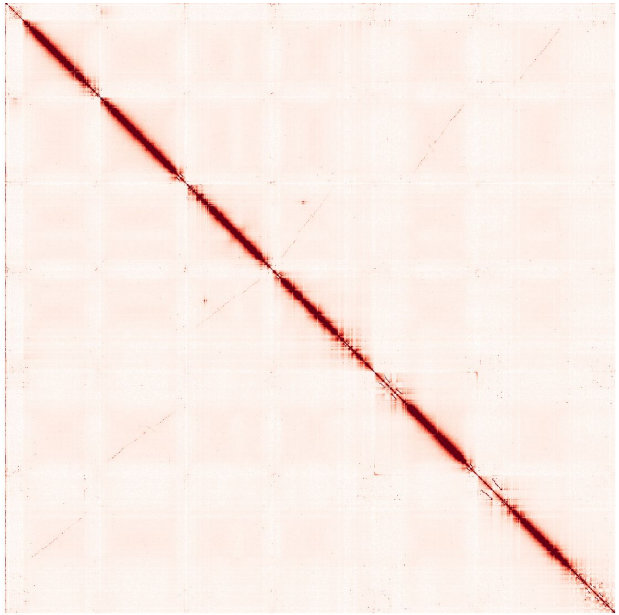




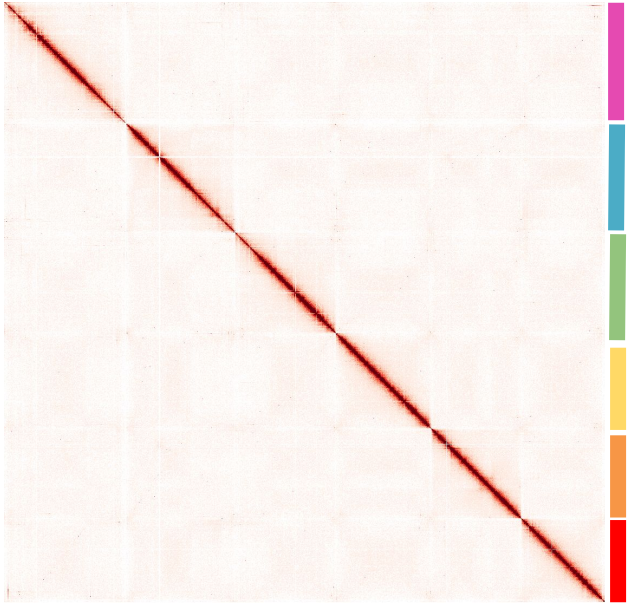
Hi-C contact map

Adineta vaga

hicstuff



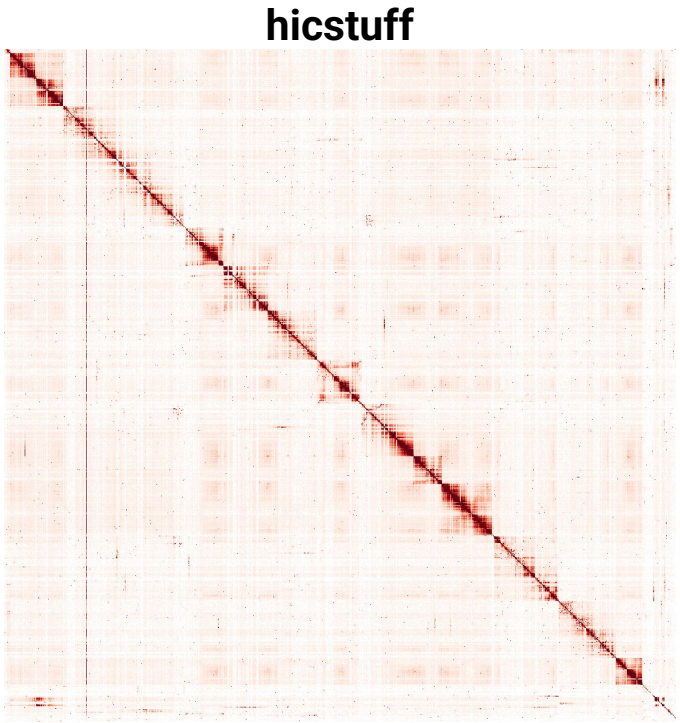
problematic assembly



improved assembly

Hi-C contact map

Xenia sp.



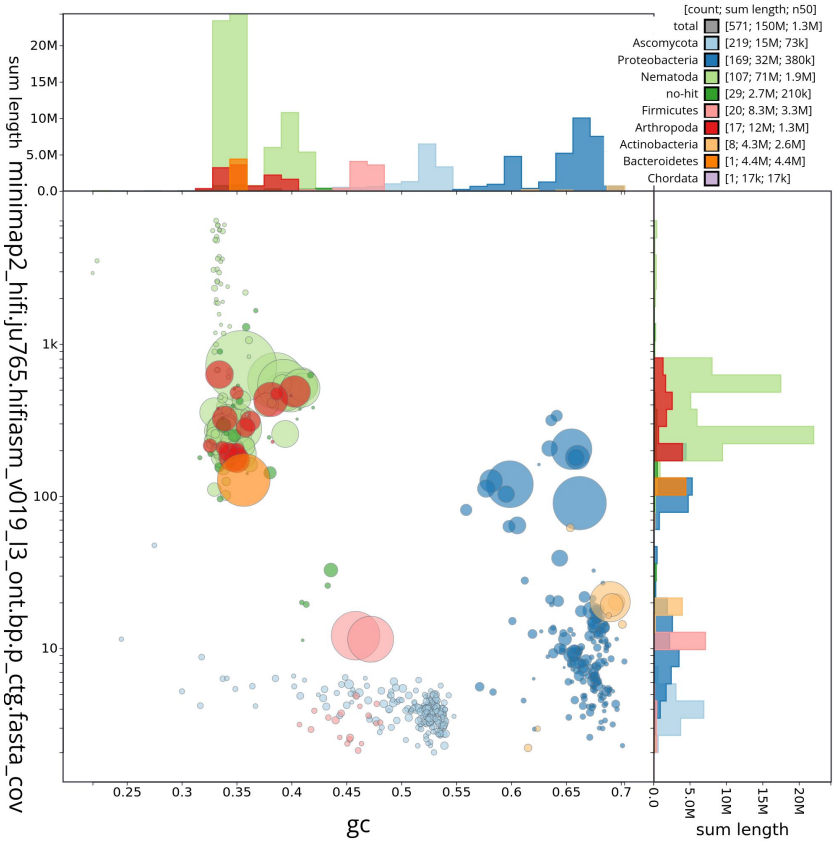
Super problematic

Contaminants

Propanagrolaimus JU765

Blobtools

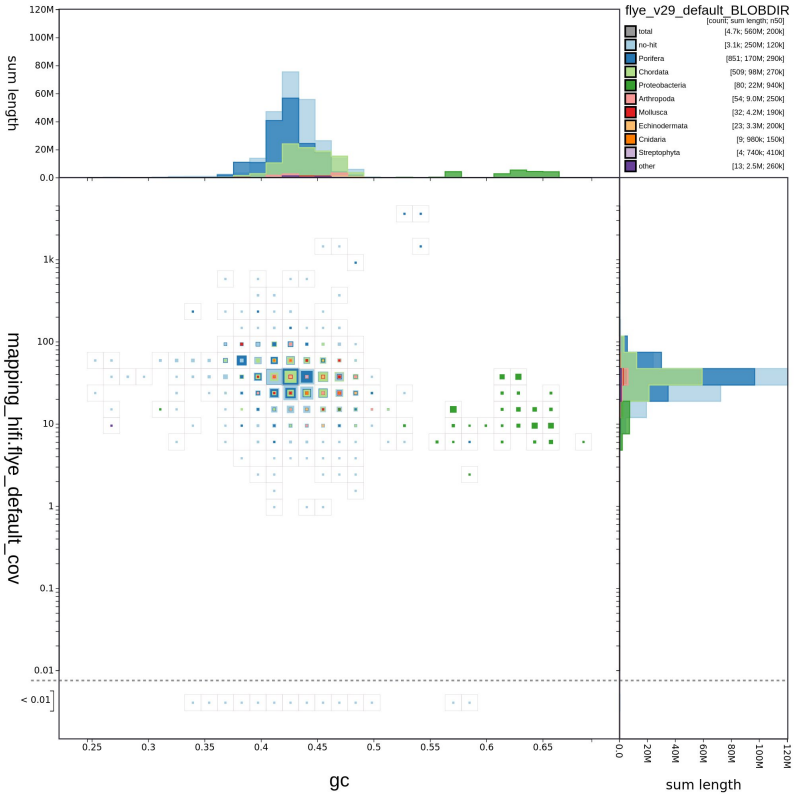
- Taxonomy based on BLAST/DIAMOND search
- Orthologs (BUSCO)
- Coverage depth



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