ECES 450 Tutorial 8

SALEH REFAHI AND DREW MANGIONE

What is metagenomic assembly?

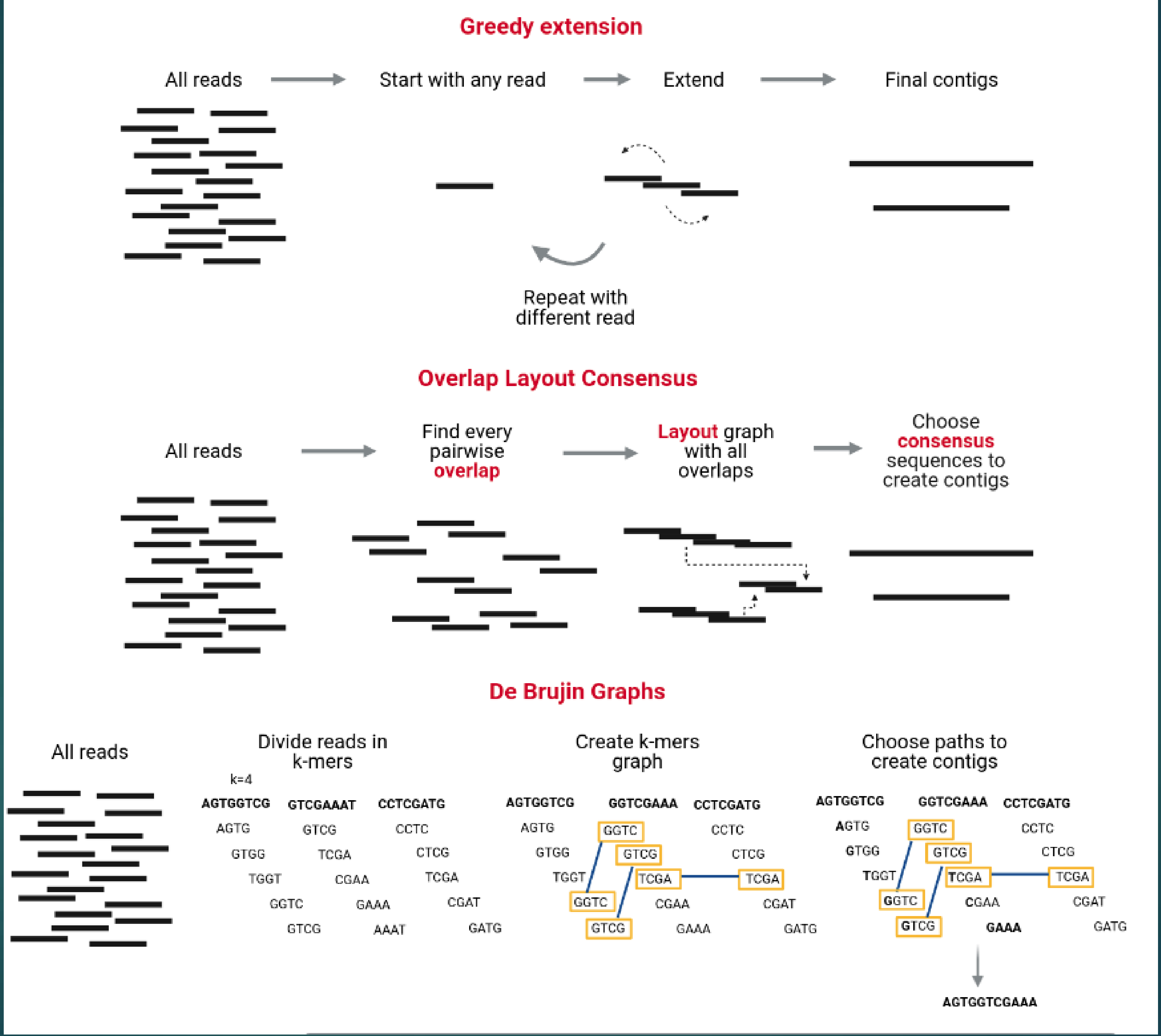
- Metagenomics is the extraction, sequencing, and analysis of the combined genomic DNA from an entire microbiome sample
- Samples will have DNA from many different organisms
- Need to reconstruct the genomes of the various organisms

How metagenomic assembly works

- Looks for reads that have overlapping segments
- Reads are combined to create contigs
- Multiple combinations might work together, needs to find best match
- Contigs are strung together into scaffolds

Types of assembly methods

- Greedy extension
  - Simplest method, computationally efficient
  - Randomly selects read and finds other reads that overlap
  - Can result in suboptimal assemblies
- Overlap Layout Consensus
  - Finds every pair of reads that of overlap
  - Lays all pairs out in graph structure, and generates a consensus merging pairs
  - Good for long-read sequences, but computationally demanding
- De Bruijn graphs
  - Creates every possible k-mer for each read
  - Finds reads with the identical k-mers and links them together
  - Very computationally efficient for large sets of short reads



Greedy extension

All reads → Start with any read → Extend → Final contigs

Repeat with different read

Overlap Layout Consensus

All reads → Find every pairwise **overlap** → **Layout** graph with all overlaps → Choose **consensus** sequences to create contigs

De Brujin Graphs

All reads

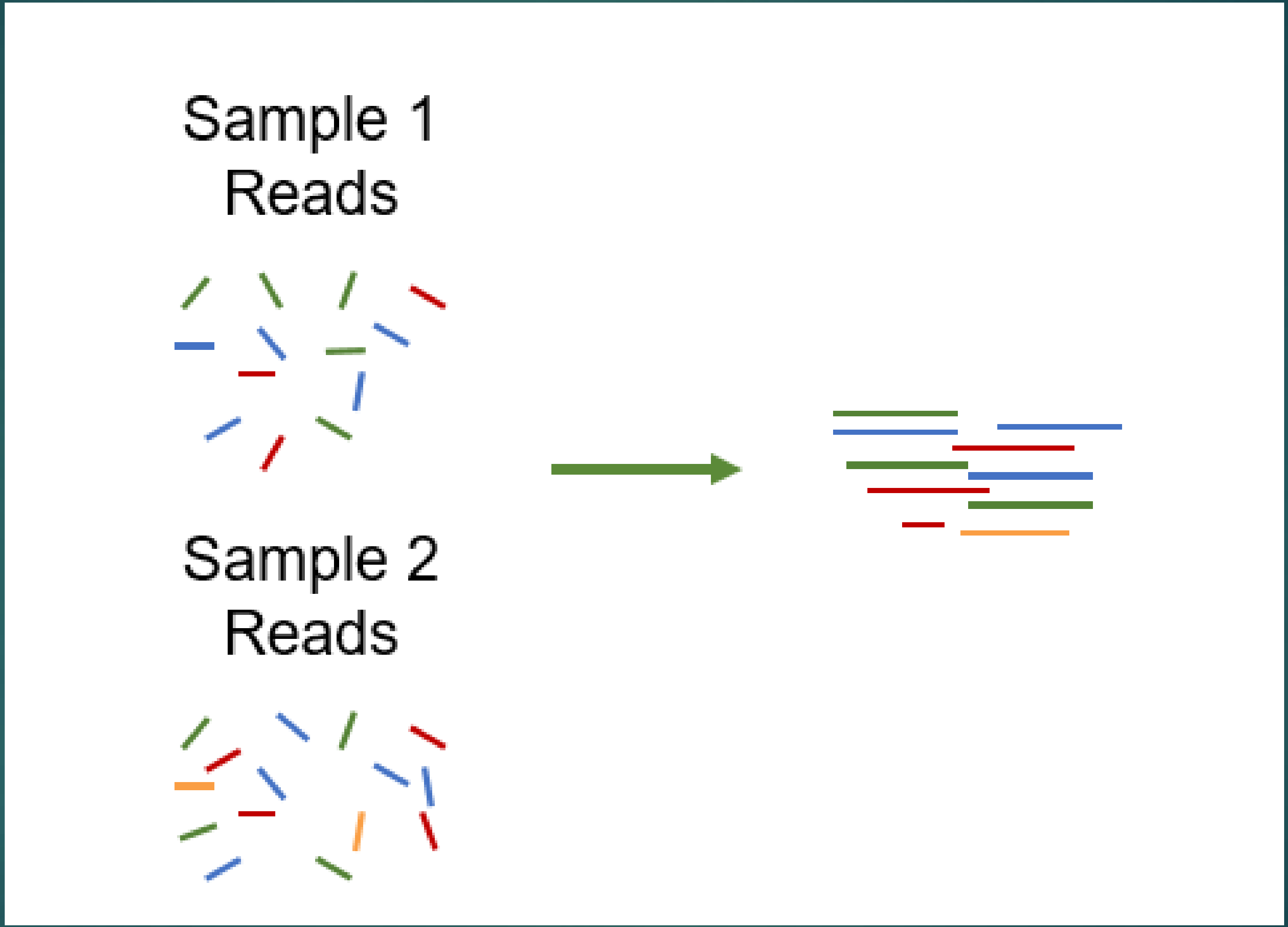
Divide reads in k-mers

Create k-mers graph

Choose paths to create contigs

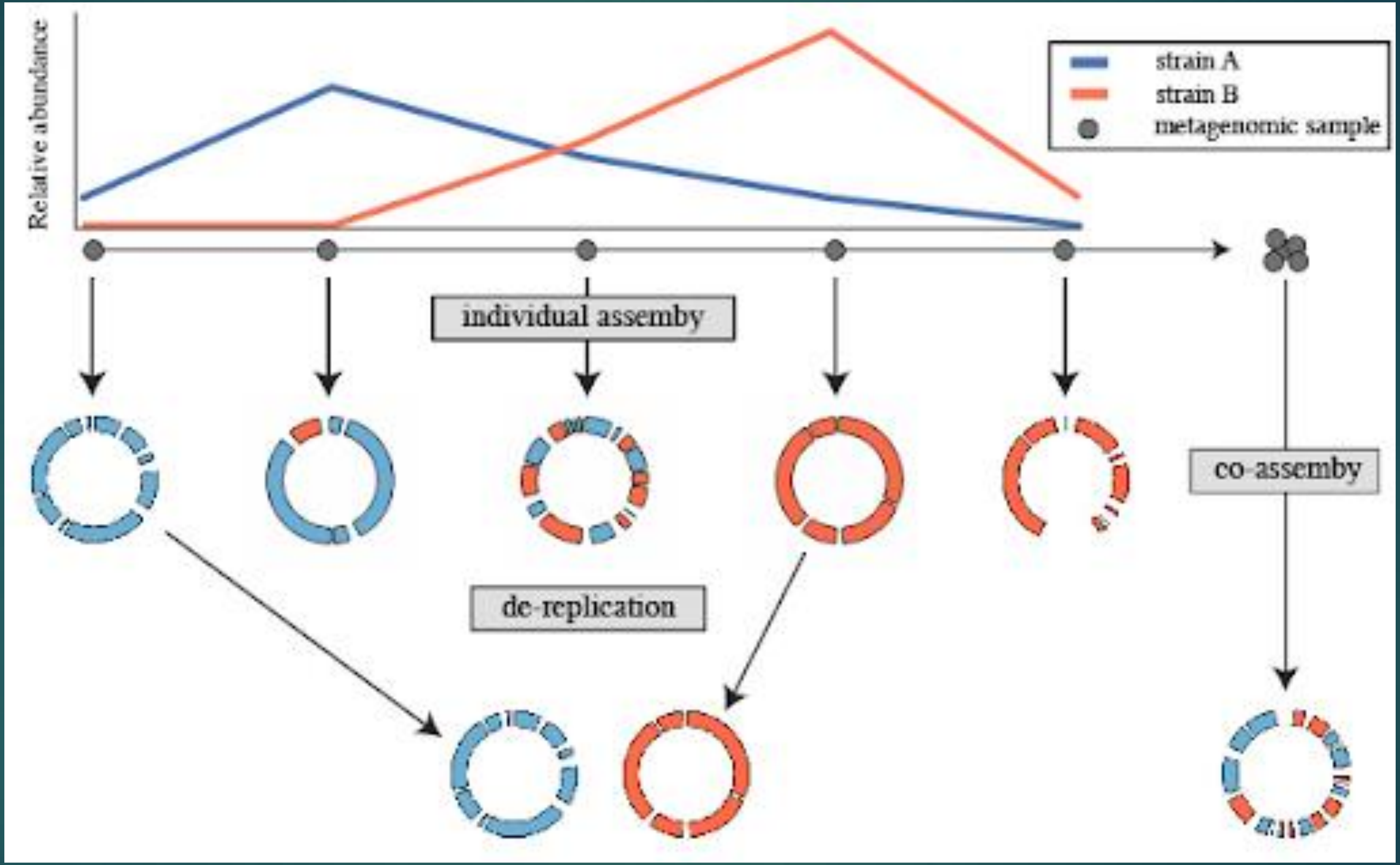
Individual vs co-assembly

- Individual assembly sequences the reads from each sample independently
  - Each sample might have different contigs from the same genome
  - De-replication can be performed to combine the results of all the individual assemblies
  - Allows for assemblies to be specifically tailored to the conditions of each sample
- Co-assembly sequences the reads from all samples at once
  - Reads are combined into a single pool for making contigs
  - Results in more complete sequences due to larger dataset
  - Helps sequence lower abundant organisms



Sample 1 Reads

Sample 2 Reads


Relative abundance
strain A
strain B
metagenomic sample
individual assemby
de-replication
co-assemby

Galaxy
PROJECT

What is Galaxy?

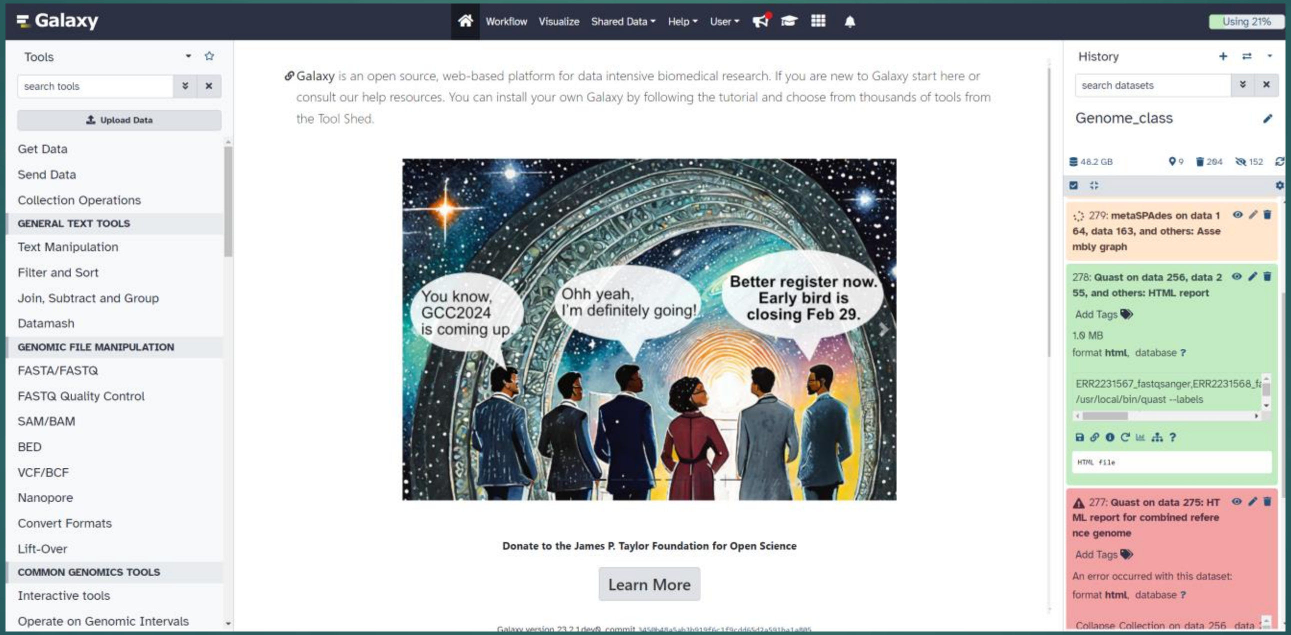
- Developed by researchers at Penn State, John Hopkins University, and Oregon Health & Science University
- Open-source program that aims to make computational biology available to those without computer expertise
- Can run as a local program or hosted as a webserver
- Serves as a platform for scheduling tasks using a wide array of tools
- Various organizations run free, publicly-available Galaxy servers
  - Each Galaxy can have different sets of tools depending on the owner's field of research

Using Galaxy for Metagenomic assembly

- MEGAhit
  - Single node assembler used for assembling large and complex genomes using de Bruijin graph method
  - Very computationally efficient
- MetaSPAdes (Meta St. Petersburg Genome Assembler)
  - Assembler specifically built for metagenomic assembly, also using de Bruijin graph method
  - Can result in better contigs, but is computationally intensive
- QUAST (Quality Assessment Tool for Genome Assemblies)
  - Gives an overview on the quality of a genome assembly, using various metrics
  - Can give insight into which assembler and settings produce better results for a given dataset

Galaxy Exploration: Interactive Session!

- https://usegalaxy.org/u/saleh_refahi/h/genomeclass



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Donate to the James P. Taylor Foundation for Open Science

Galaxy version 23.2.1.dev0, commit 3450b48a5ab3b919f6c1f9cdd65d2a591ba1a805

Genome_class

48.2 GB 9 204 152

279: **metaSPAdes on data 164, data 163, and others: Assembly graph**

278: **Quast on data 256, data 255, and others: HTML report**

1.0 MB

format **html**, database **?**

```
ERR2231567_fastqsanger,ERR2231568_f
/usr/local/bin/quast --labels
```

```
HTML file
```

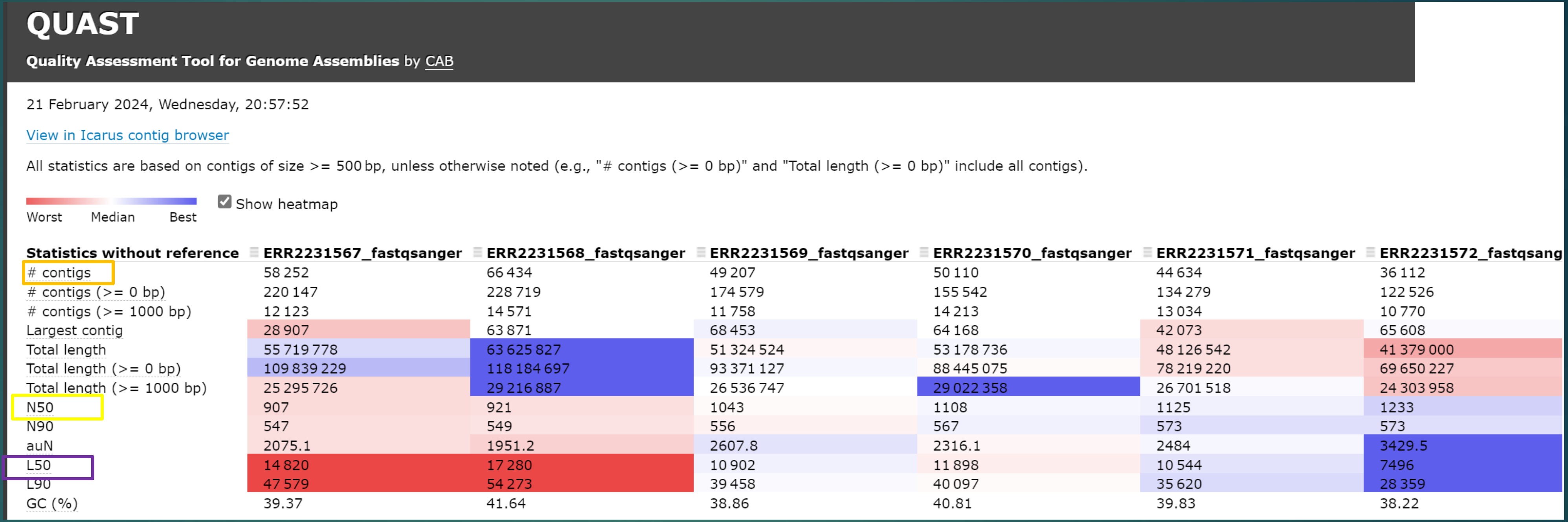
277: **Quast on data 275: HTML report for combined reference genome**

An error occurred with this dataset:

format **html**, database **?**

Collapse Collection on data 256, data

Quast: Statistics without Reference



QUAST

Quality Assessment Tool for Genome Assemblies by CAB

21 February 2024, Wednesday, 20:57:52

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

☑ Show heatmap

Statistics without reference	ERR2231567_fastqsanger	ERR2231568_fastqsanger	ERR2231569_fastqsanger	ERR2231570_fastqsanger	ERR2231571_fastqsanger	ERR2231572_fastqsang
# contigs	58 252	66 434	49 207	50 110	44 634	36 112
# contigs (>= 0 bp)	220 147	228 719	174 579	155 542	134 279	122 526
# contigs (>= 1000 bp)	12 123	14 571	11 758	14 213	13 034	10 770
Largest contig	28 907	63 871	68 453	64 168	42 073	65 608
Total length	55 719 778	63 625 827	51 324 524	53 178 736	48 126 542	41 379 000
Total length (>= 0 bp)	109 839 229	118 184 697	93 371 127	88 445 075	78 219 220	69 650 227
Total length (>= 1000 bp)	25 295 726	29 216 887	26 536 747	29 022 358	26 701 518	24 303 958
N50	907	921	1043	1108	1125	1233
N90	547	549	556	567	573	573
auN	2075.1	1951.2	2607.8	2316.1	2484	3429.5
L50	14 820	17 280	10 902	11 898	10 544	7496
L90	47 579	54 273	39 458	40 097	35 620	28 359
GC (%)	39.37	41.64	38.86	40.81	39.83	38.22

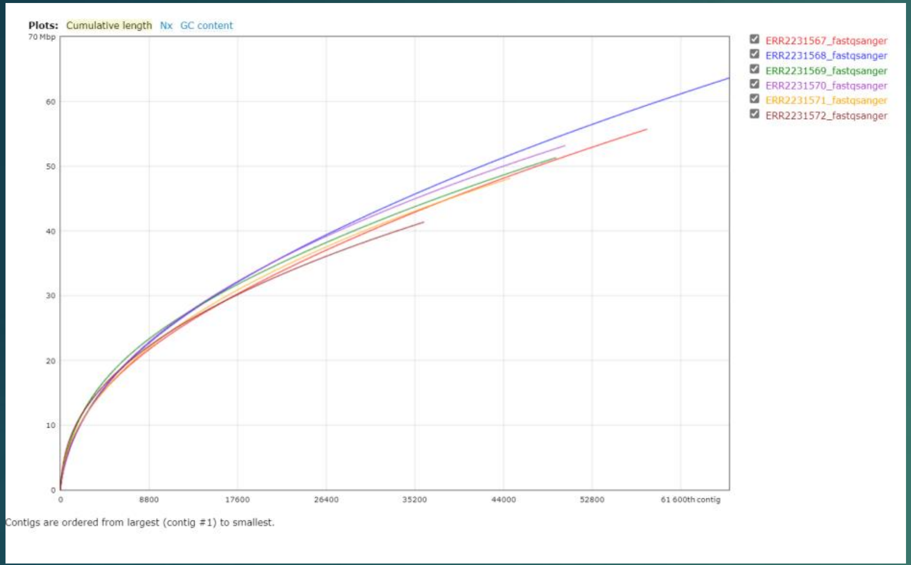
contigs longer than 500bp

L50: number of contigs equal to or longer than N50

N50: is measure of contiguity; length for which the collection of all contigs of that length or longer covers at least half an assembly.
For Quiz ! Let's consider 9 contigs with the lengths 2, 3, 4, 5, 6, 7, 8, 9, and 12:

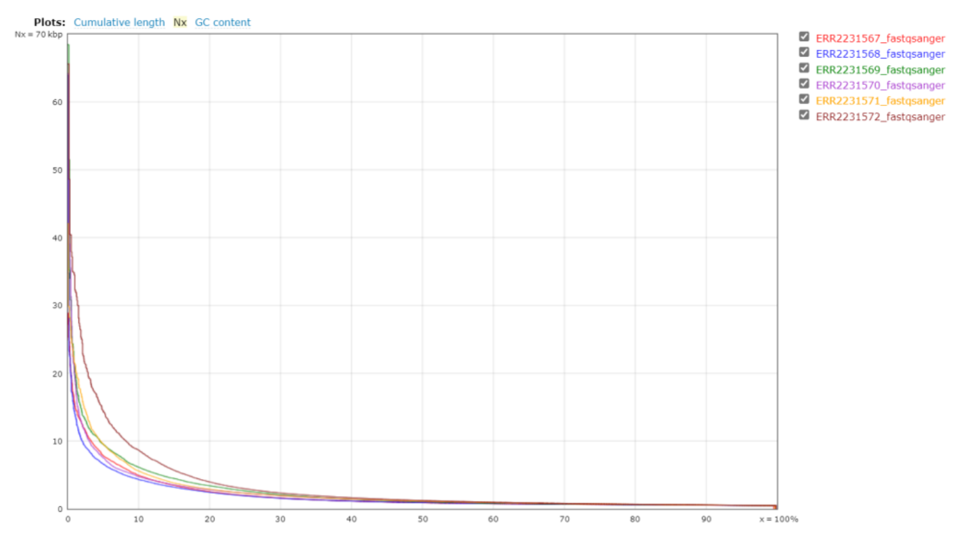
- The sum of the length is 56
- Half of the sum is 28
- 12 + 9 + 8 = 28 (half the length of the sequence)
- N50 = 8 ; L50 =3

Quast: Statistics without Reference



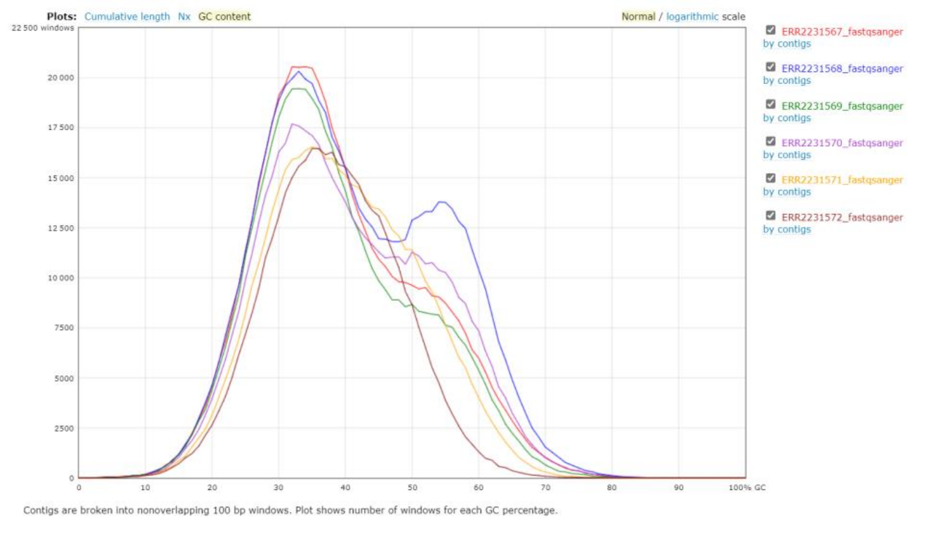
Contigs are ordered from largest (contig #1) to smallest.

Cumulative length



Nx

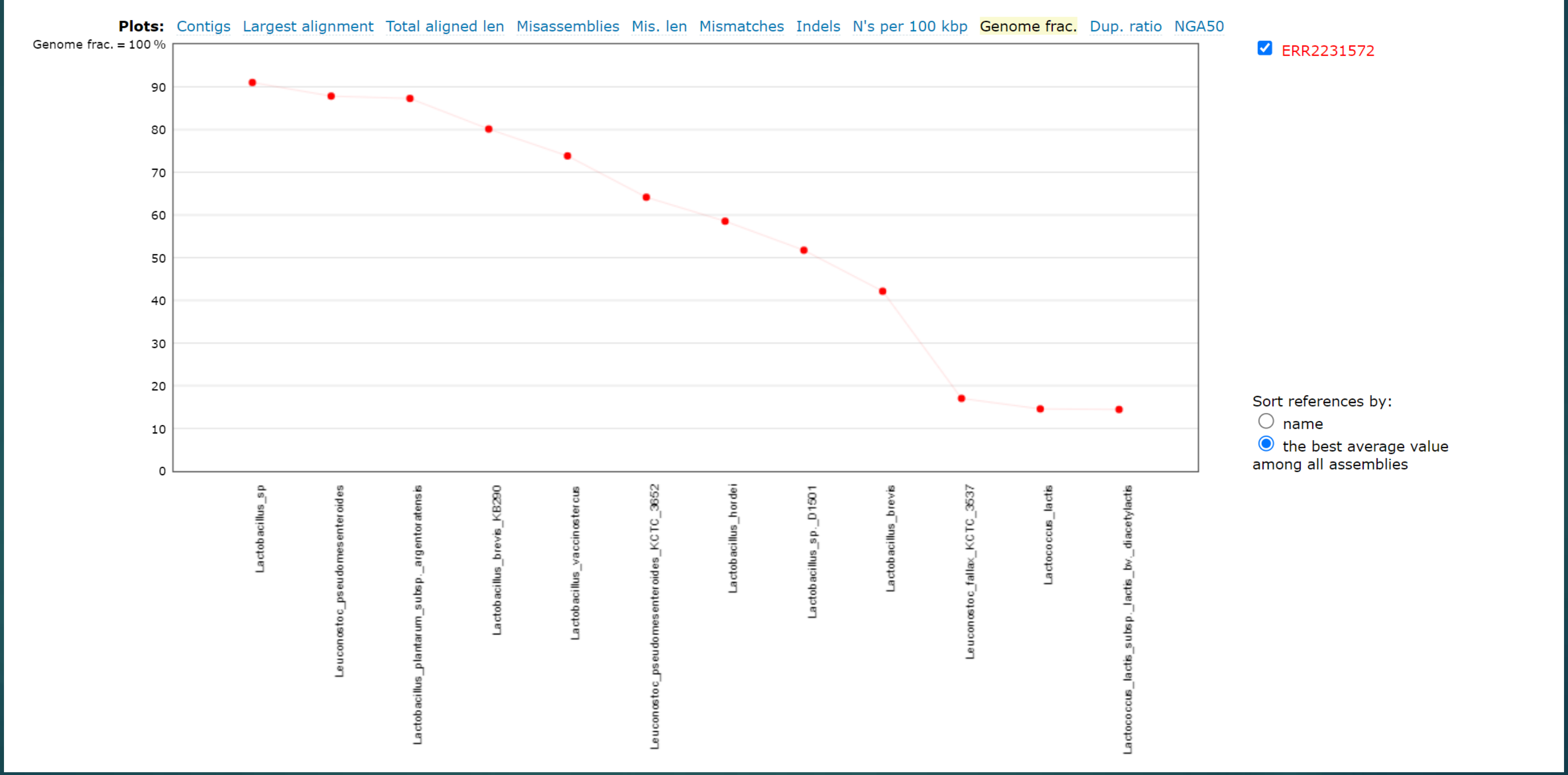
GC Content



Contigs are broken into nonoverlapping 100 bp windows. Plot shows number of windows for each GC percentage.

Quast: Statistics with Reference

- **Genome fraction (%)**: percentage of aligned bases in the reference genome (Silva)

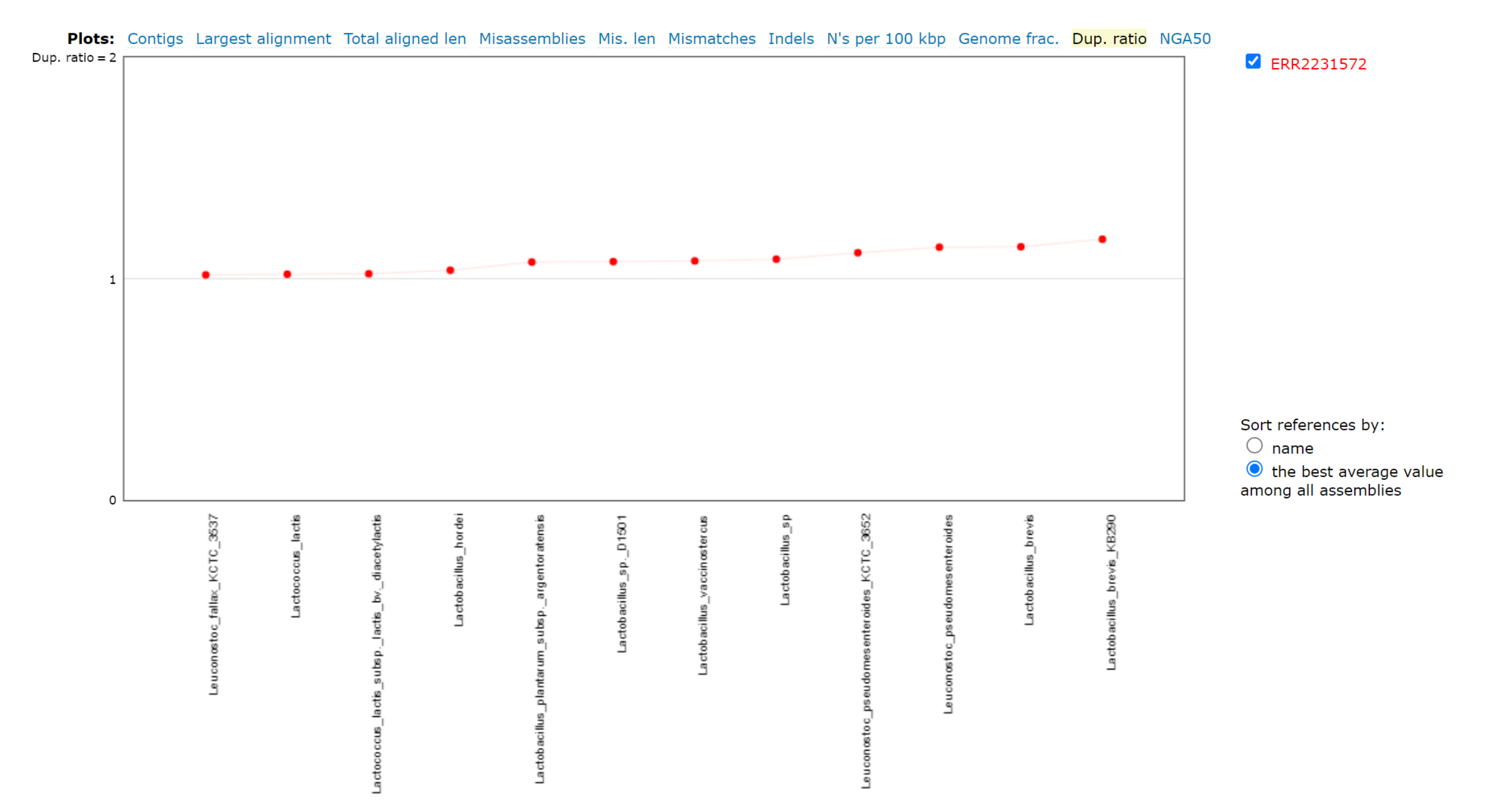


☑ ERR2231572

Sort references by:
○ name
● the best average value among all assemblies

Quast: Statistics with Reference

- **Duplication ratio**: total number of aligned bases / genome fraction * reference length
- If an assembly contains many contigs that cover the same regions of the reference, the duplication ratio may be much larger than 1.


Plots:
Contigs
Largest alignment
Total aligned len
Misassemblies
Mis. len
Mismatches
Indels
N's per 100 kbp
Genome frac.
Dup. ratio
NGA50
Dup. ratio = 2
1
0
Leuconostoc_fallax_KCTC_3537
Lactococcus_lactis
Lactococcus_lactis_subsp._lactis_bv._diacetylactis
Lactobacillus_hordei
Lactobacillus_plantarum_subsp._argentoratensis
Lactobacillus_sp._D1501
Lactobacillus_vaccinostercus
Lactobacillus_sp
Leuconostoc_pseudomesenteroides_KCTC_3652
Leuconostoc_pseudomesenteroides
Lactobacillus_brevis
Lactobacillus_brevis_KB290

☑ ERR2231572

Sort references by:

○ name

● the best average value among all assemblies

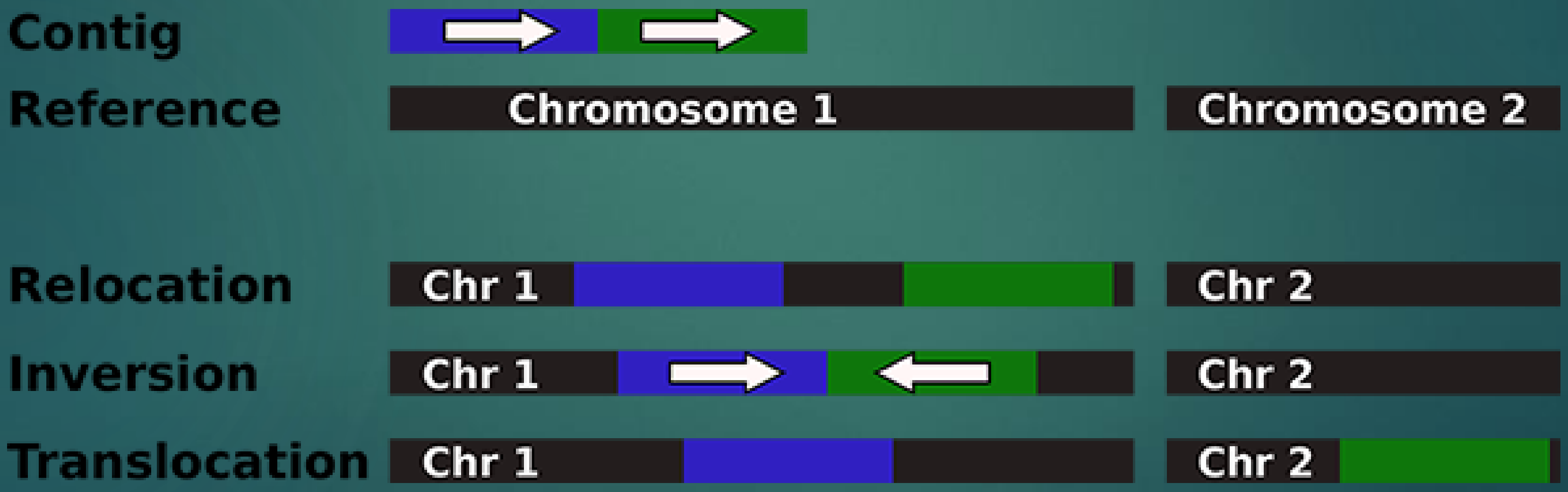
Misassemblies

- **Misassemblies**: joining sequences that should not be adjacent.

Relocation occur based on signal from two mappings of the same contig against the same chromosome which are separated by an unmapped region of at least 1kbp (or overlapped by 1kbp)

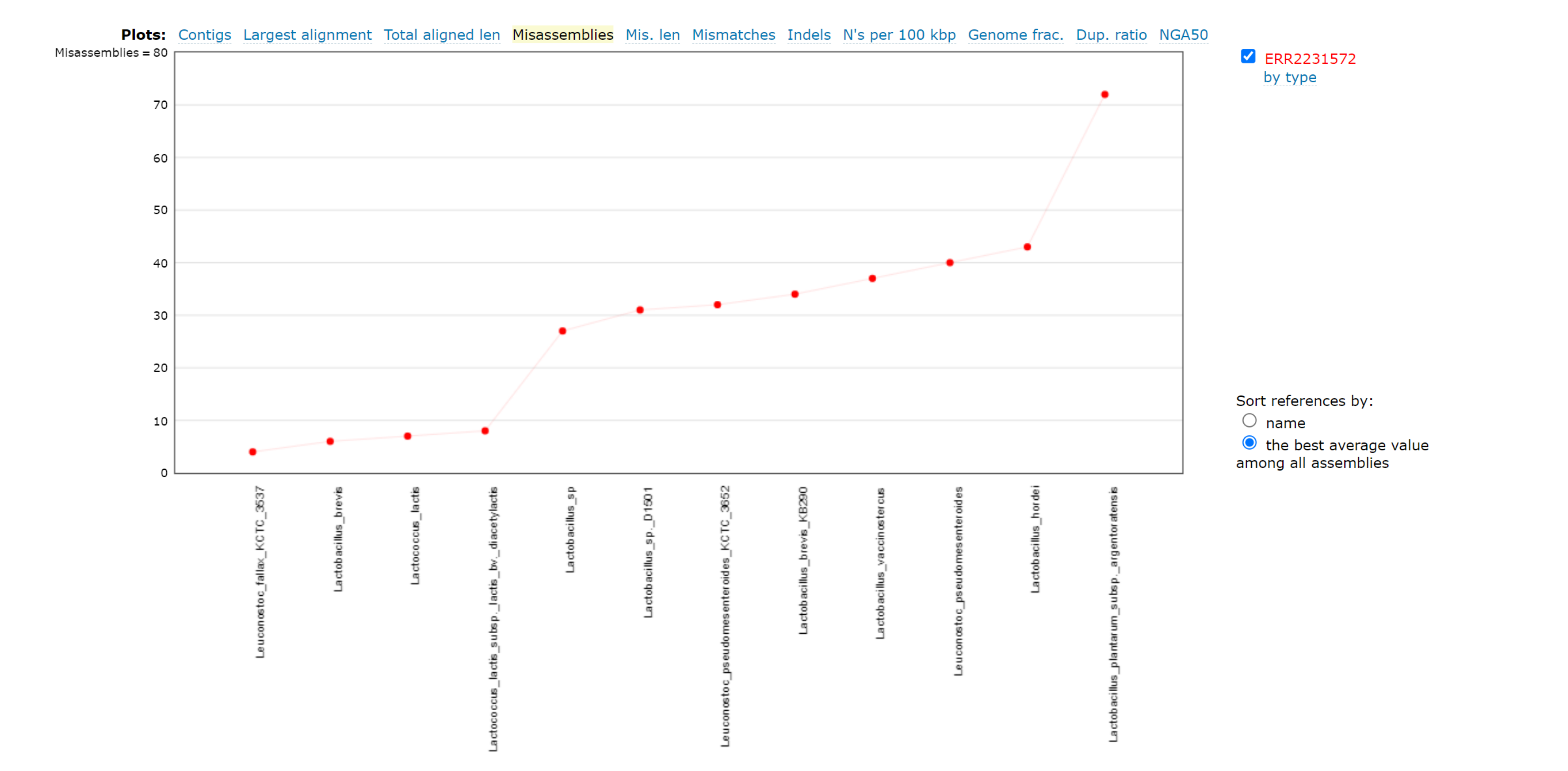
Translocation occur when a contig has mapped on more than one reference chromosomes

Inversion occurs when a contig has two consecutive mappings on the same chromosome but in different strands



Quast: Statistics with Reference

- Quast identifies missassemblies by mapping the contigs to the reference genomes


Plots: Contigs Largest alignment Total aligned len Misassemblies Mis. len Mismatches Indels N's per 100 kbp Genome frac. Dup. ratio NGA50
Misassemblies = 80
70
60
50
40
30
20
10
0
Leuconostoc_fallax_KCTC_3537
Lactobacillus_brevis
Lactococcus_lactis
Lactococcus_lactis_subsp._lactis_bv._diacetylactis
Lactobacillus_sp
Lactobacillus_sp._D1501
Leuconostoc_pseudomesenteroides_KCTC_3652
Lactobacillus_brevis_KB290
Lactobacillus_vaccinostercus
Leuconostoc_pseudomesenteroides
Lactobacillus_hordei
Lactobacillus_plantarum_subsp._argentoratensis
by type

☑ ERR2231572

Sort references by:
○ name
● the best average value among all assemblies

Sources

Polunina, Polina, and Bérénice Batut. “Assembly / Hands-on: Assembly of Metagenomic Sequencing Data.” Galaxy Training Network, Galaxy Training Network, 21 Feb. 2024, training.galaxyproject.org/training-material/topics/assembly/tutorials/metagenomics-assembly/tutorial.html.

Afgan, Enis et al. “The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.” Nucleic acids research vol. 44,W1 (2016): W3-W10. doi:10.1093/nar/gkw343

Assembling a Metagenome and Recovering “Genomes” with Anvi’o, astrobiomike.github.io/metagenomics/metagen_anvio#:~:text=%E2%80%9CCo%2Dassembly%E2%80%9D%20refers%20to,reads%20from%20that%20individual%20sample. Accessed 20 Feb. 2024.

Ghurye, Jay S et al. “Metagenomic Assembly: Overview, Challenges and Applications.” The Yale journal of biology and medicine vol. 89,3 353-362. 30 Sep. 2016