SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Assembly Validation

elixir

excelerate

Assembly

SciLifeLab

- What is an assembly?

"Draft de novo genome assemblies are now available for many organisms. These **assemblies are point estimates** of the true genome sequences. **Each is a specific hypothesis**, drawn from among many alternative hypotheses, of the sequence of a genome. Assembly uncertainty, the inability to distinguish between multiple alternative assembly hypotheses, can be due to real variation between copies of the genome in the sample, errors and ambiguities in the sequenced data and assumptions and heuristics of the assemblers. Most assemblers select a single assembly according to ad hoc criteria, and do not yet report and quantify the uncertainty of their outputs. Those assemblers that do report uncertainty take different approaches to describing multiple assembly hypotheses and the support for each."

Howison et al. 2013. Bioinformatics.

Evaluating assemblies

SciLifeLab

- What can we do to evaluate an assembly?
  - Polish first (Pilon, NanoPolish, Quiver/Arrow, etc).
  - Assembly statistics
  - K-mer statistics
  - Assembly graph structure
  - Read alignment statistics and properties
  - Contamination assessment
  - Gene space statistics
  - Comparative alignment

Basic assembly statistics

SciLifeLab

Percentage of assembly in scaffolded contigs	4.2%	
Percentage of assembly in unscaffolded contigs	95.8%	
Average number of contigs per scaffold	1.0	
Average length of break (>25 Ns) between contigs in scaffold	191	
Number of contigs	9082	
Number of contigs in scaffolds	72	
Number of contigs not in scaffolds	9010	
Total size of contigs	22857451	
Longest contig	621740	
Shortest contig	56	
Number of contigs > 1K nt	2527	27.8%
Number of contigs > 10K nt	329	3.6%
Number of contigs > 100K nt	34	0.4%
Number of contigs > 1M nt	0	0.0%
Number of contigs > 10M nt	0	0.0%
Mean contig size	2517	
Median contig size	571	
N50 contig length	25795	
L50 contig count	158	
NG50 contig length	188047	
LG50 contig count	8	
N50 contig - NG50 contig length difference	162252	
contig %A	28.57	
contig %C	21.46	
contig %G	21.39	
contig %T	28.58	
contig %N	0.01	
contig %non-ACGTN	0.00	
Number of contig non-ACGTN nt	0	

Basic assembly statistics

SciLifeLab

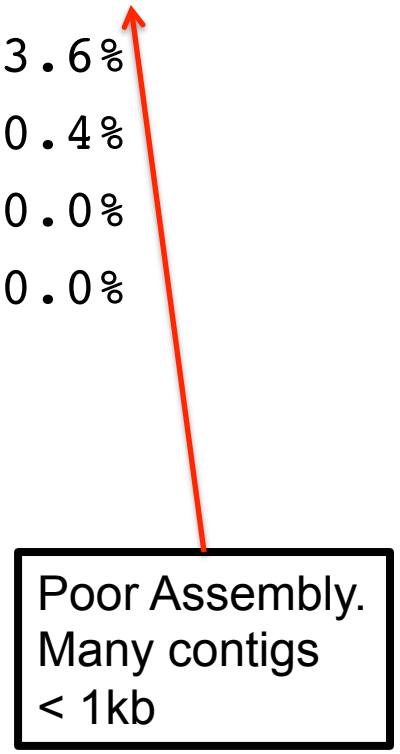
Percentage of assembly in scaffolded contigs	4.2%
Percentage of assembly in unscaffolded contigs	95.8%
Average number of contigs per scaffold	1.0
Average length of break (>25 Ns) between contigs in scaffold	191
Number of contigs	9082
Number of contigs in scaffolds	72
Number of contigs not in scaffolds	9010
Total size of contigs	22857451

Estimated genome size **4.66Mb** (4660000)

Basic assembly statistics

SciLifeLab

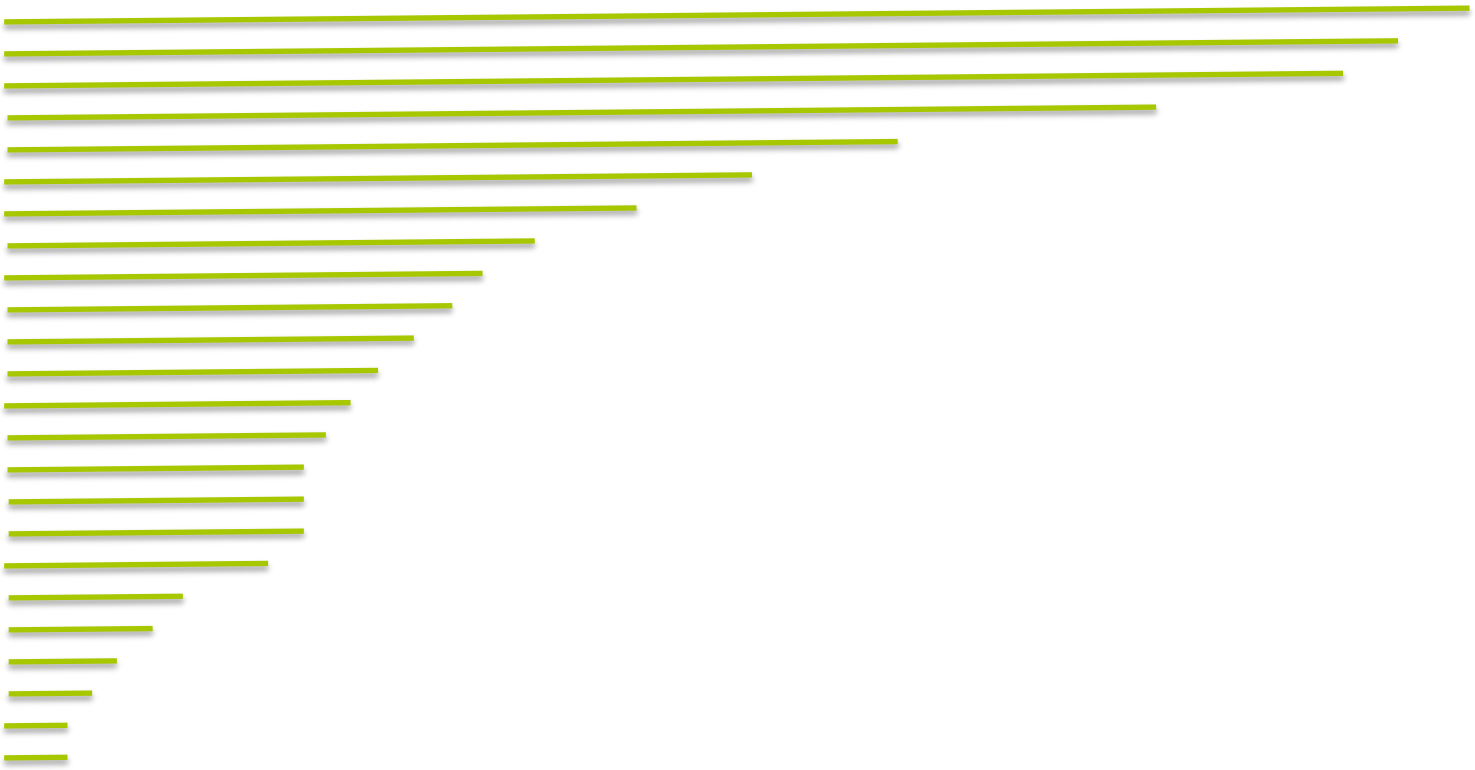
Longest contig	621740	
Shortest contig	56	
Number of contigs > 1K nt	2527	27.8%
Number of contigs > 10K nt	329	3.6%
Number of contigs > 100K nt	34	0.4%
Number of contigs > 1M nt	0	0.0%
Number of contigs > 10M nt	0	0.0%
Mean contig size	2517	
Median contig size	571	
N50 contig length	25795	
L50 contig count	158	
NG50 contig length	188047	
LG50 contig count	8	



Basic assembly statistics

SciLifeLab

- N50 is a common statistical measure of sequence length.
  - The size of the smallest contig in the set of largest contigs that make up 50% of the assembly size.



Basic assembly statistics

SciLifeLab

- N50 has multiple disadvantages though
  - N50 is not a measure of assembly correctness
    - It only measures sequence contiguity
  - N50 is not meaningful for different assembly sizes.
    - It's not comparable across species, and technically even the same genome.
  - N50 does not improve for near complete assemblies.
    - Once you have good scaffolds, only small contigs remain.
  - N50 is biased if short sequences are excluded.
    - Often shorter contigs are filtered out from the assembly.

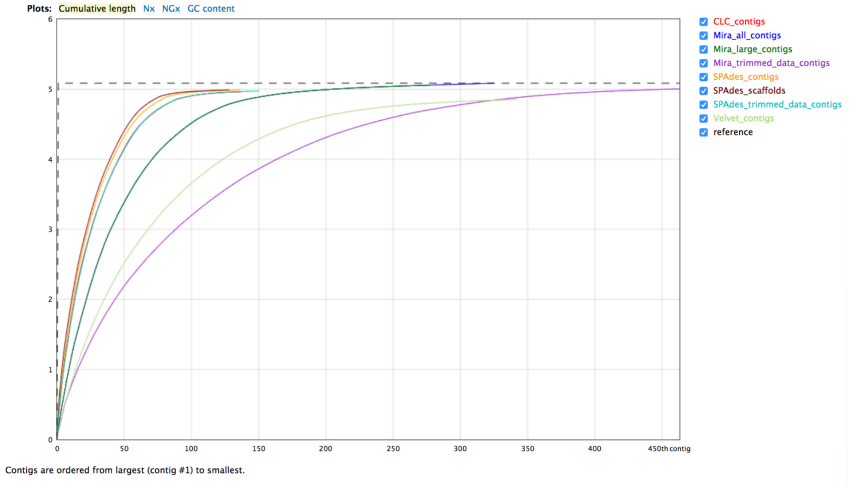
Basic assembly statistics

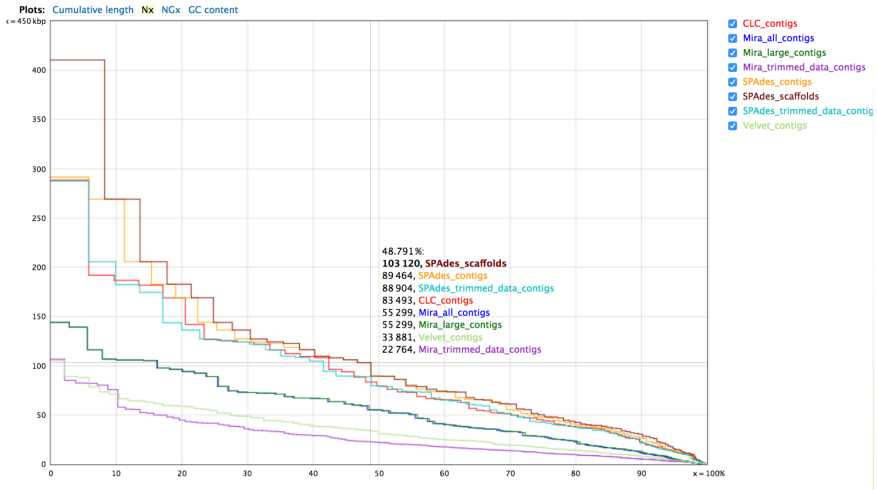
SciLifeLab

- A better statistic is NG50
  - The size of the smallest contig in the set of largest contigs that make up 50% of the (estimated) **genome** size (not assembly).
    - It is still only a measure of sequence contiguity, but comparable for the same genome.
    - There is still a limit on when it will not improve further.
    - Smaller contigs can be filtered out without affecting the value.

Basic assembly statistics

- Tool: Quast
  - Produces comparisons of assemblies
  - Statistics include number of contigs, N50, NG50, GC content

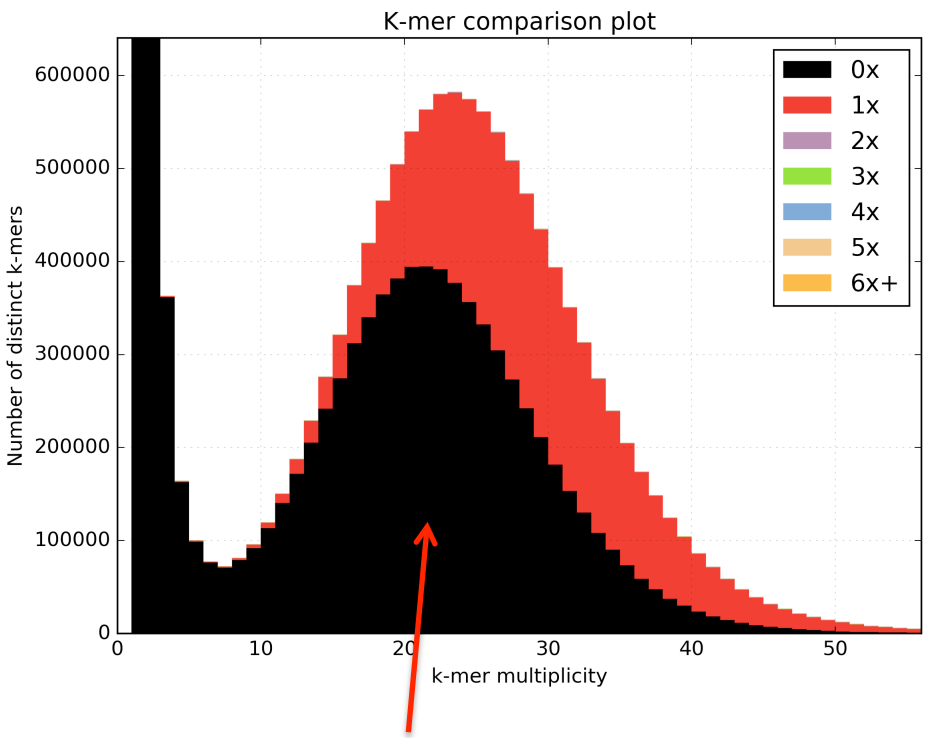




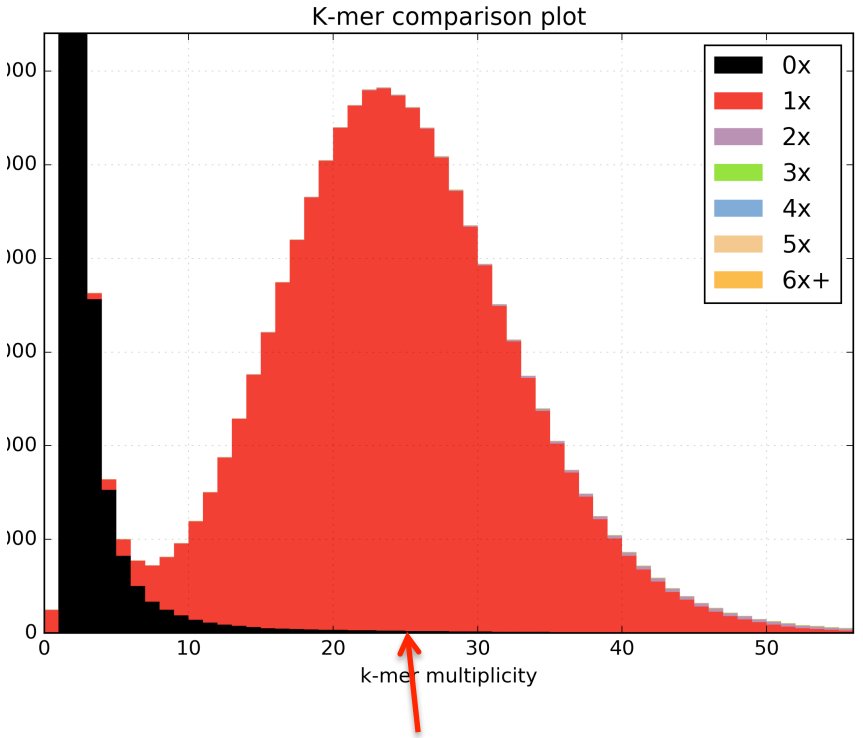
K-mer statistics

SciLifeLab

- K-mer Analysis Toolkit
  - K-mer comparison plots indicate how well the genome is assembled.



Poor - Many high frequency are k-mers missing from the assembly

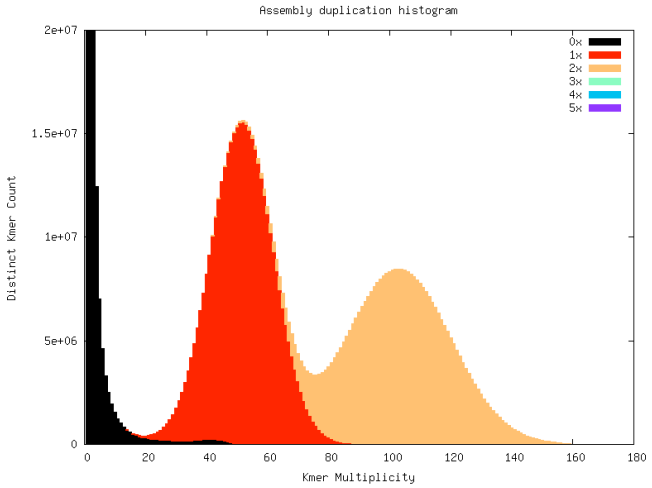


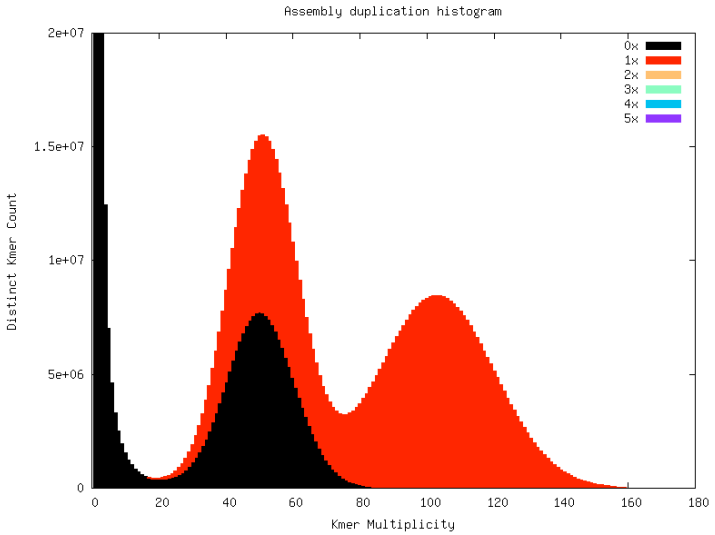
Good - Most high frequency are found in the assembly

K-mer statistics

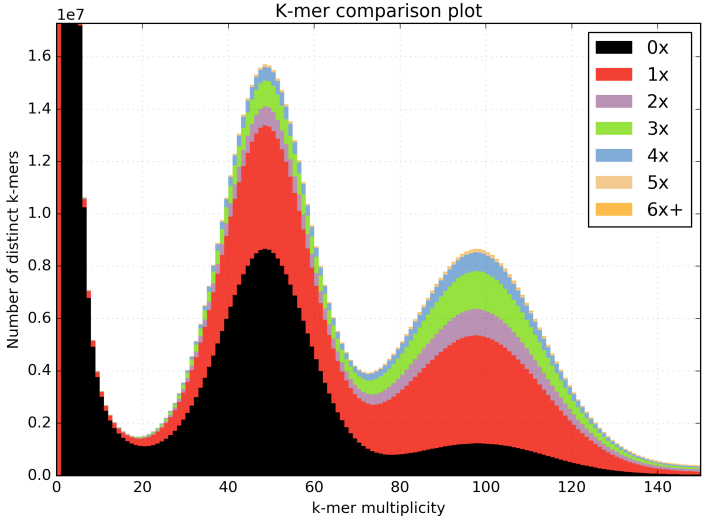
SciLifeLab

- K-mer Analysis Toolkit





Good

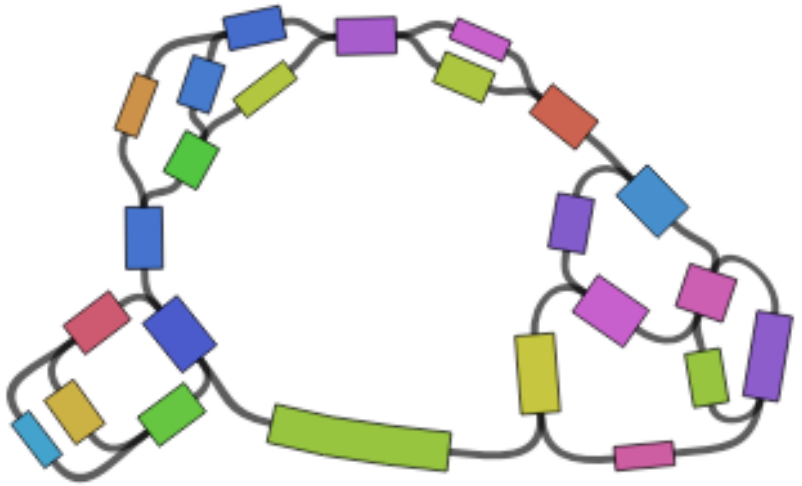


Bad

Assembly graph structure

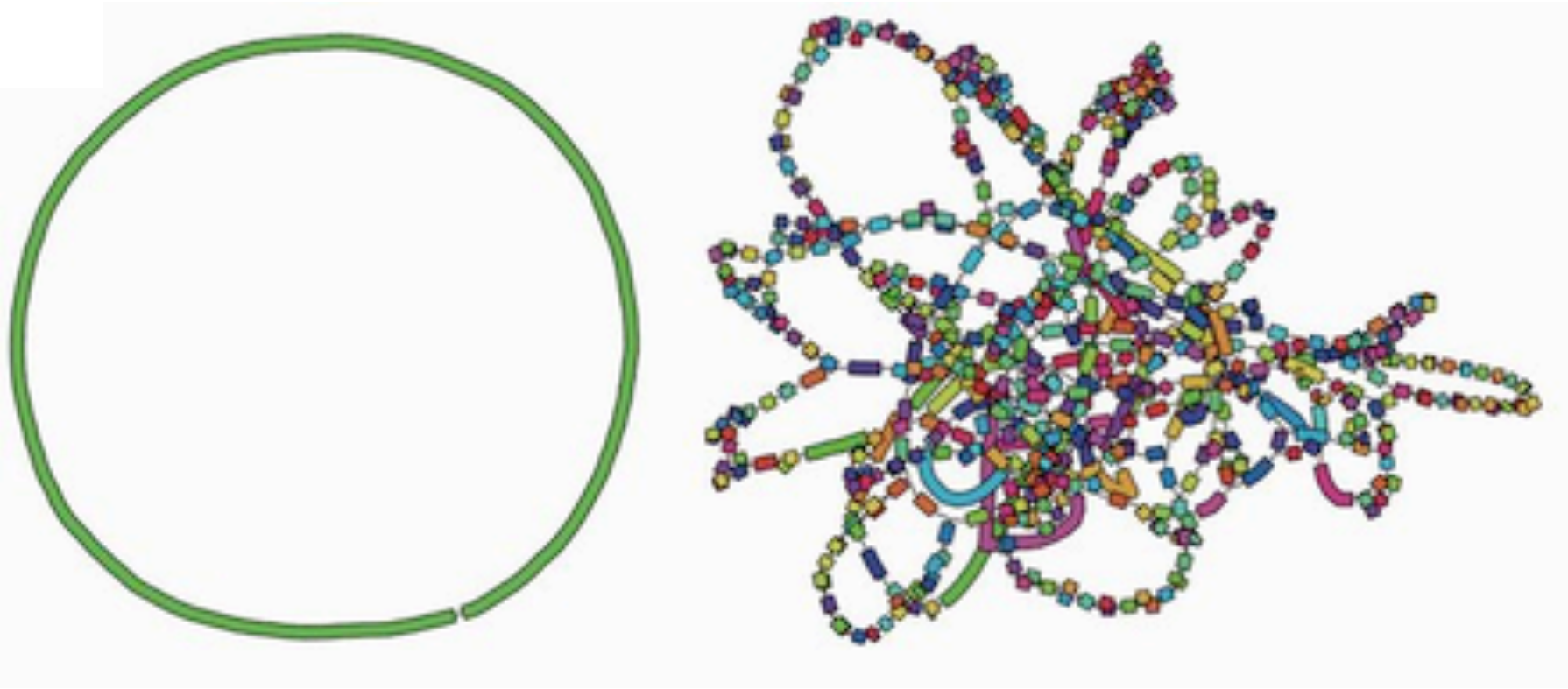
SciLifeLab

- Assemblies are written in fasta/q format.
  - Loses connection information between the contigs/scaffolds.
- Some assemblers also write GFA format (Graphical Fragment Assembly)
  - Keeps the relationship between contigs.
  - Visualized using Bandage



Assembly graph structure

SciLifeLab

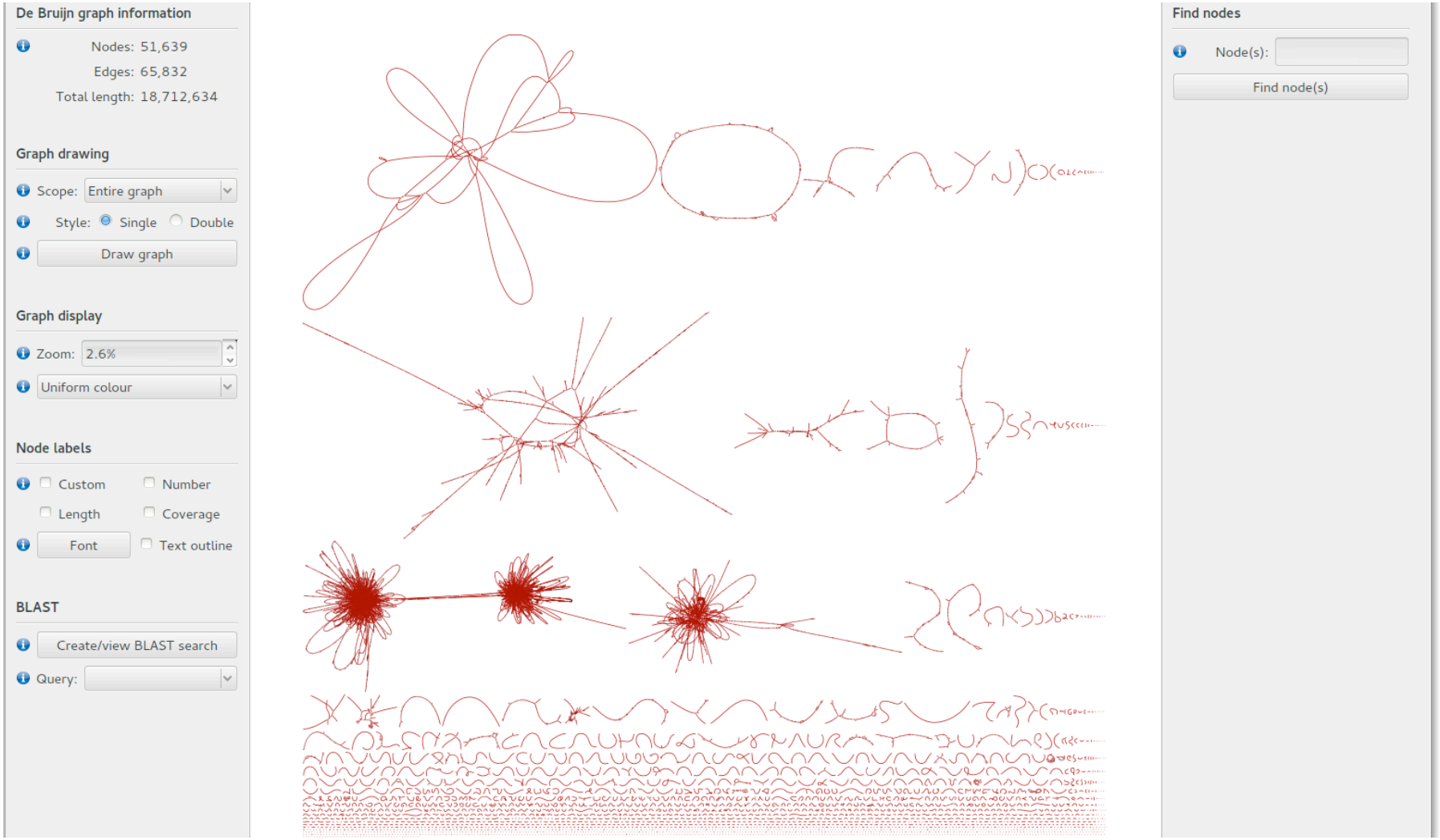


Ideal bacterial assembly graph

Poor assembly

Assembly graph structure

SciLifeLab



Read alignment statistics

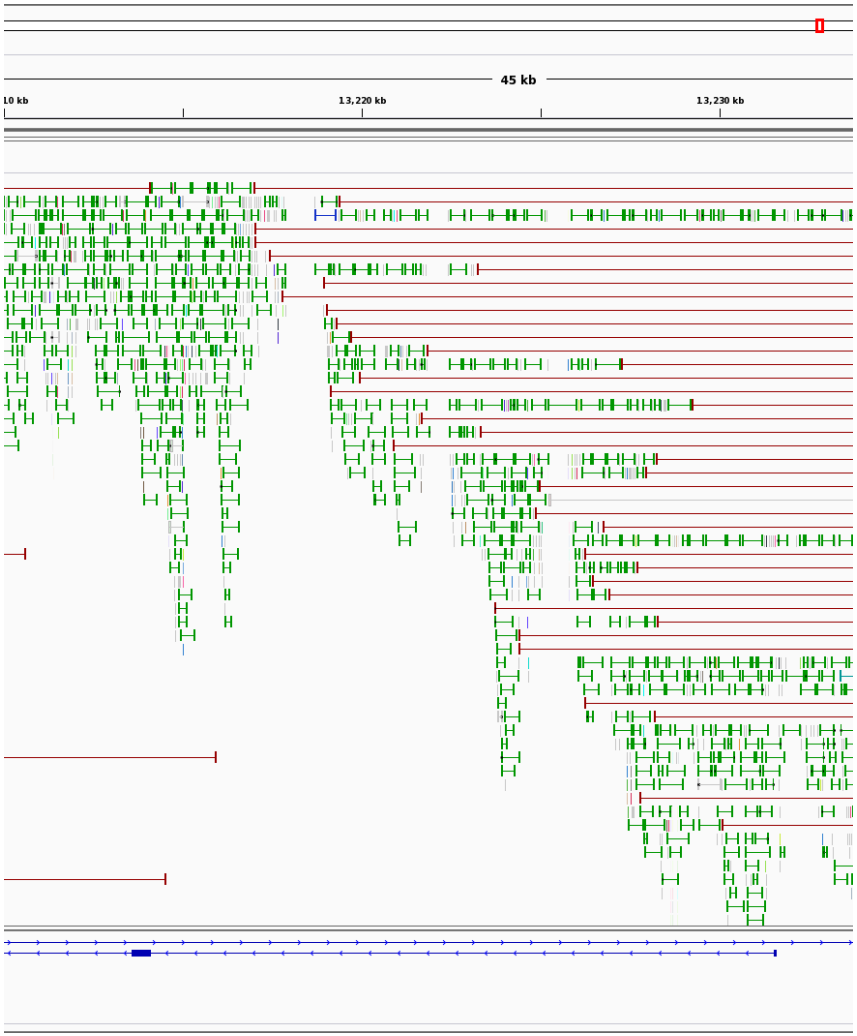
SciLifeLab

- Samtools flagstat <bamfile>

```
27190072 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
584370 + 0 supplementary
0 + 0 duplicates
25987447 + 0 mapped (95.58% : N/A)
26605702 + 0 paired in sequencing
13302851 + 0 read1
13302851 + 0 read2
23321920 + 0 properly paired (87.66% : N/A)
25250050 + 0 with itself and mate mapped
153027 + 0 singletons (0.58% : N/A)
1196126 + 0 with mate mapped to a different chr
439746 + 0 with mate mapped to a different chr (mapQ>=5)
```

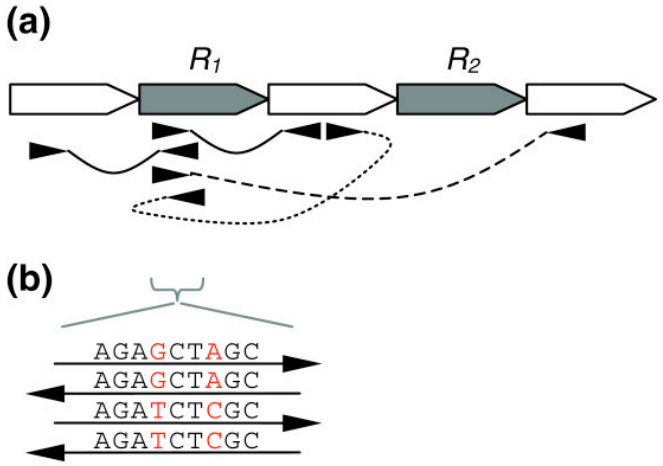
Read alignment properties

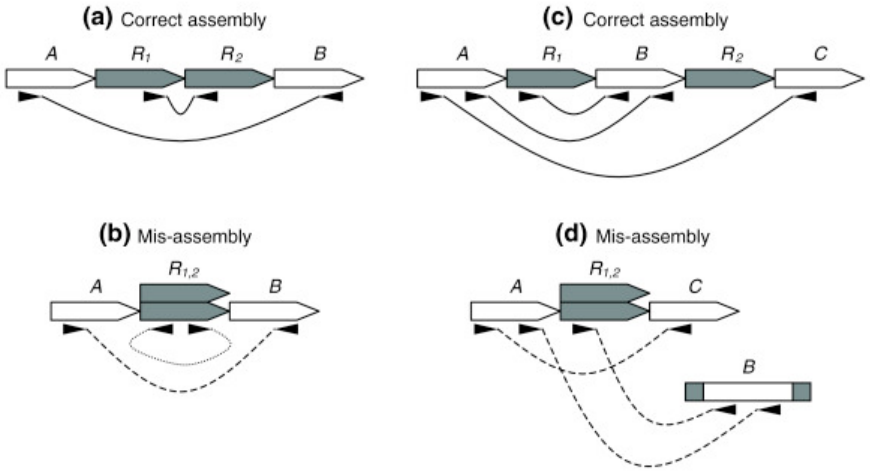
- Aligning reads back to the draft assembly tells us about data congruency.
  - Which areas of the assembly have no / reduced coverage?
  - Do paired reads align to different contigs?
  - Do paired reads align to close or too far apart?
  - Do paired reads align in the wrong orientation?

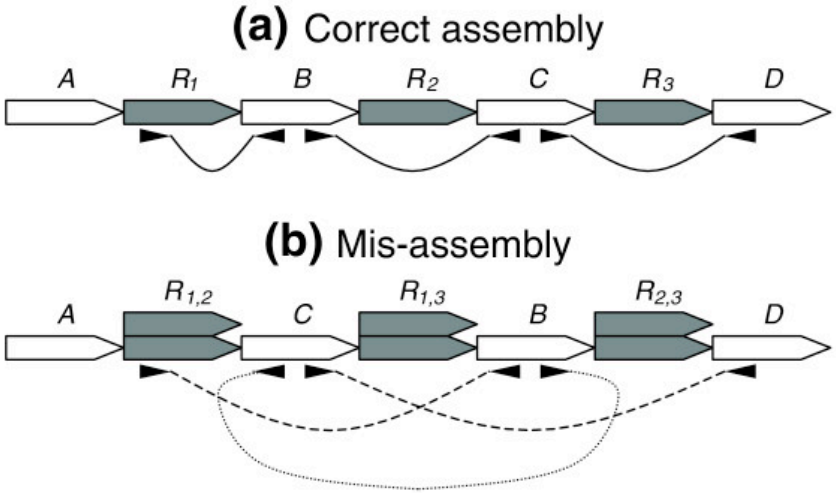


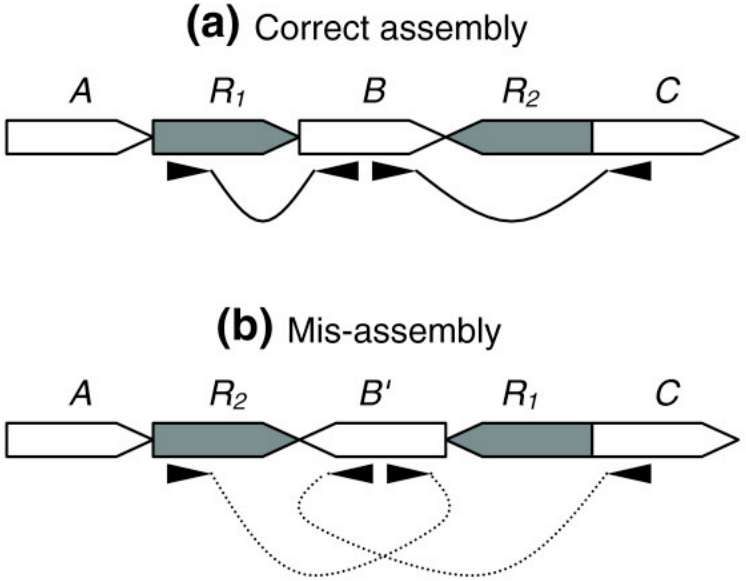
Read alignment properties

SciLifeLab









Read alignment properties

SciLifeLab

- FRCBam

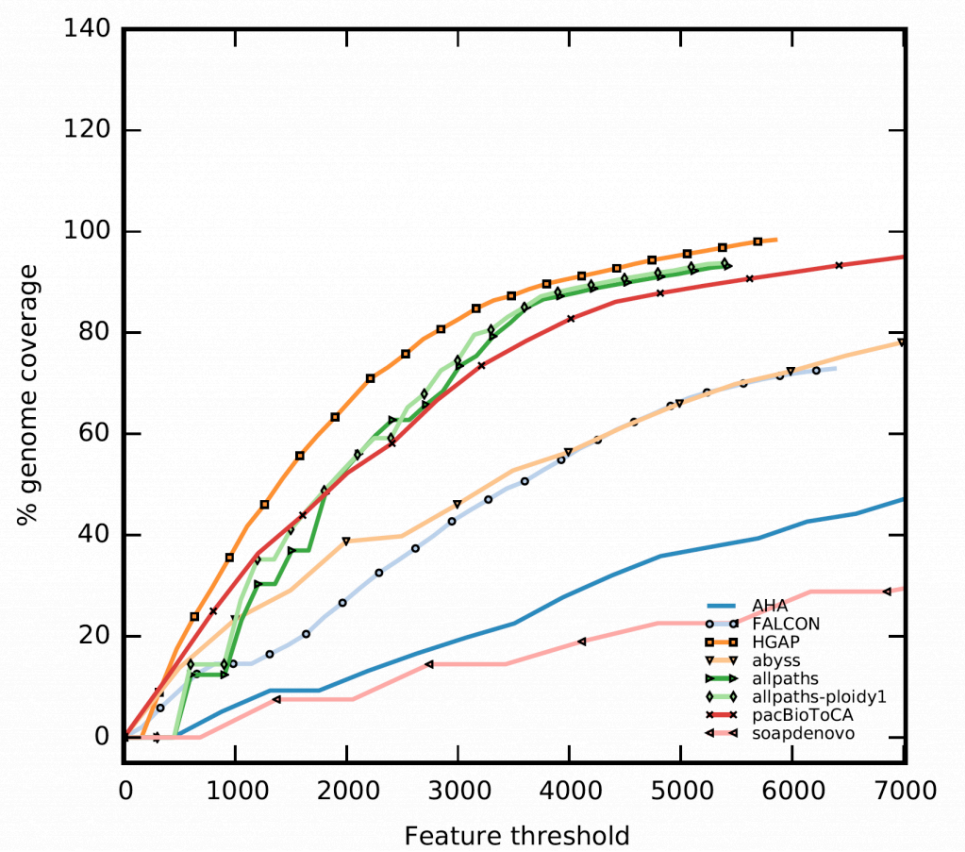
Feature	Description
LOW_COV_PE	*low read coverage* areas (all aligned reads).
HIGH_COV_PE	*high read coverage* areas (all aligned reads).
LOW_NORM_COV_PE	*low paired-read coverage* areas (only properly aligned pairs).
HIGH_NORM_COV_PE	*high paired-read coverage* areas (only properly aligned pairs).
COMPR_PE	*low CE-statistics* computed on PE-reads.
STRECH_PE	*high CE-statistics* computed on PE-reads.
HIGH_SINGLE_PE	*high number of PE reads with unmapped pair.*
HIGH_SPAN_PE	*high number of PE reads with pair mapped in a different contig/scaffold.*
HIGH_OUTIE_PE	*high number of mis-oriented or too distant PE reads.*
COMPR_MP	*low CE-statistics* computed on MP reads.
STRECH_MP	*high CE-statistics* computed on MP reads.
HIGH_SINGLE_MP	*high number of MP reads with unmapped pair.*
HIGH_SPAN_MP	*high number of MP reads with pair mapped in a different contig/scaffold.*
HIGH_OUTIE_MP	*high number of mis-oriented or too distant MP reads.*

The Table provides a brief description for each implemented feature.
doi:10.1371/journal.pone.0052210.t001

Read alignment properties

SciLifeLab

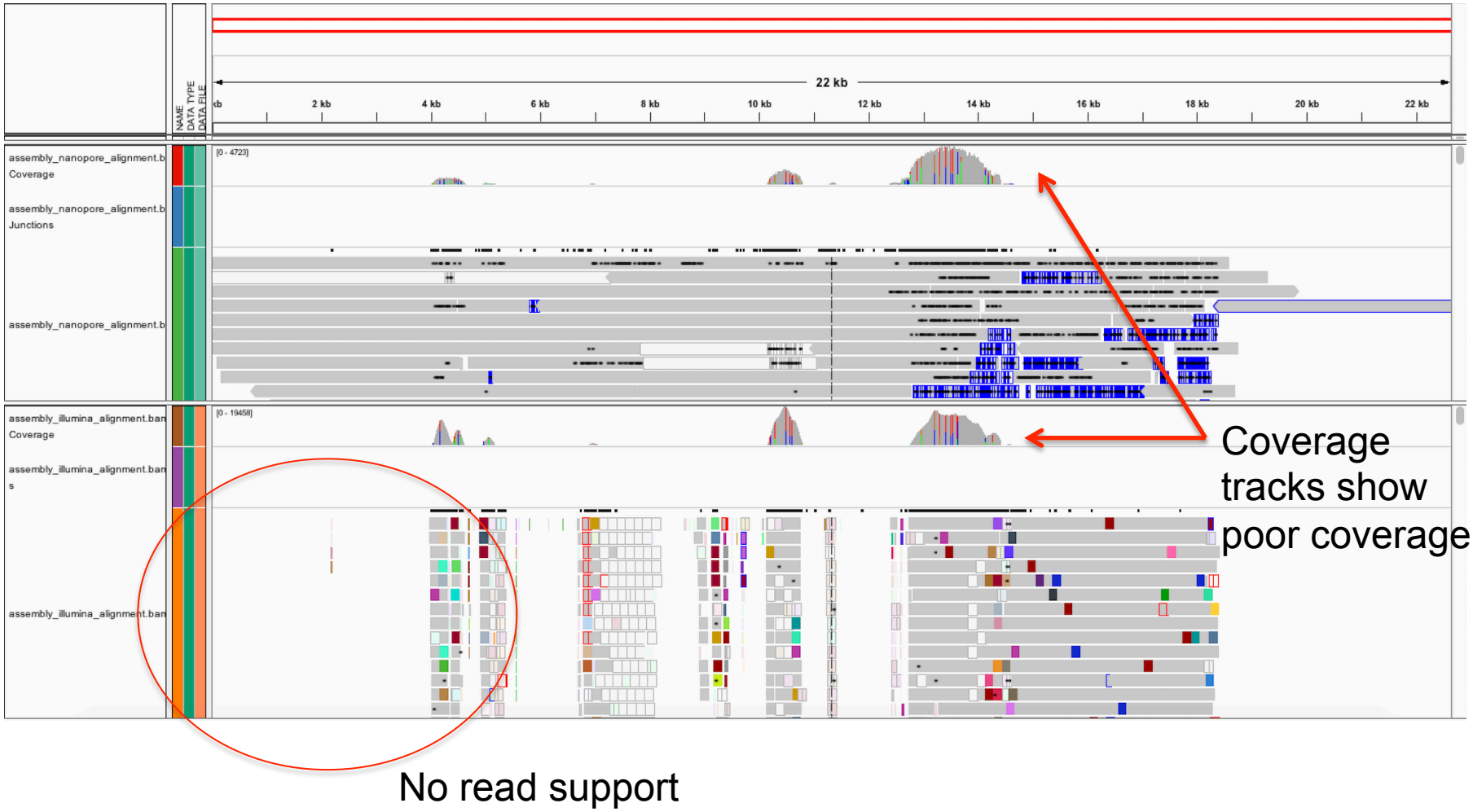
- FRCBam
  - Feature Response Curve (only comparable if estimated genome size is used).
  - The best assembly has the least features.



Read alignment properties

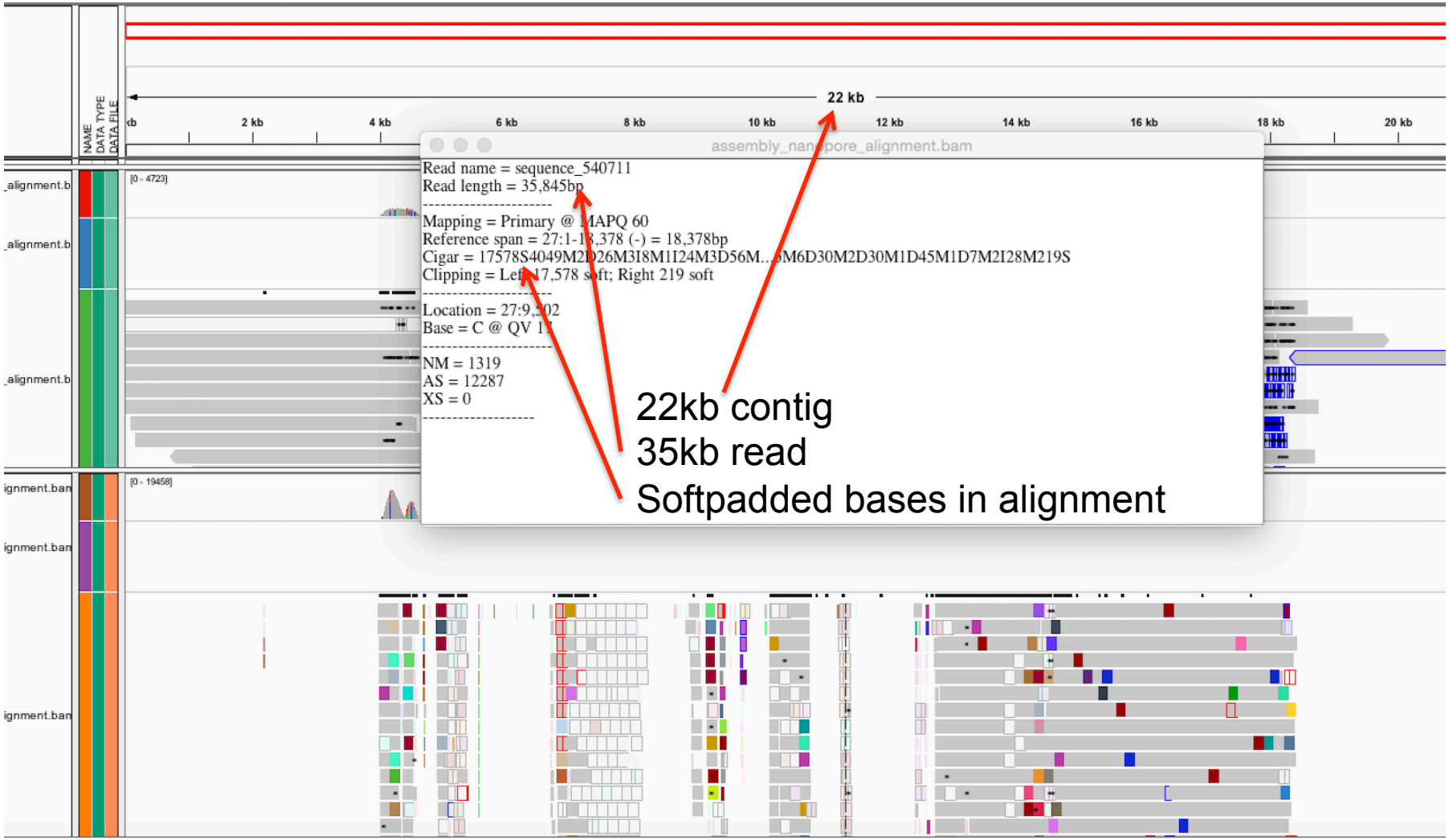
SciLifeLab

IGV - Genome Browser



Read alignment properties

SciLifeLab



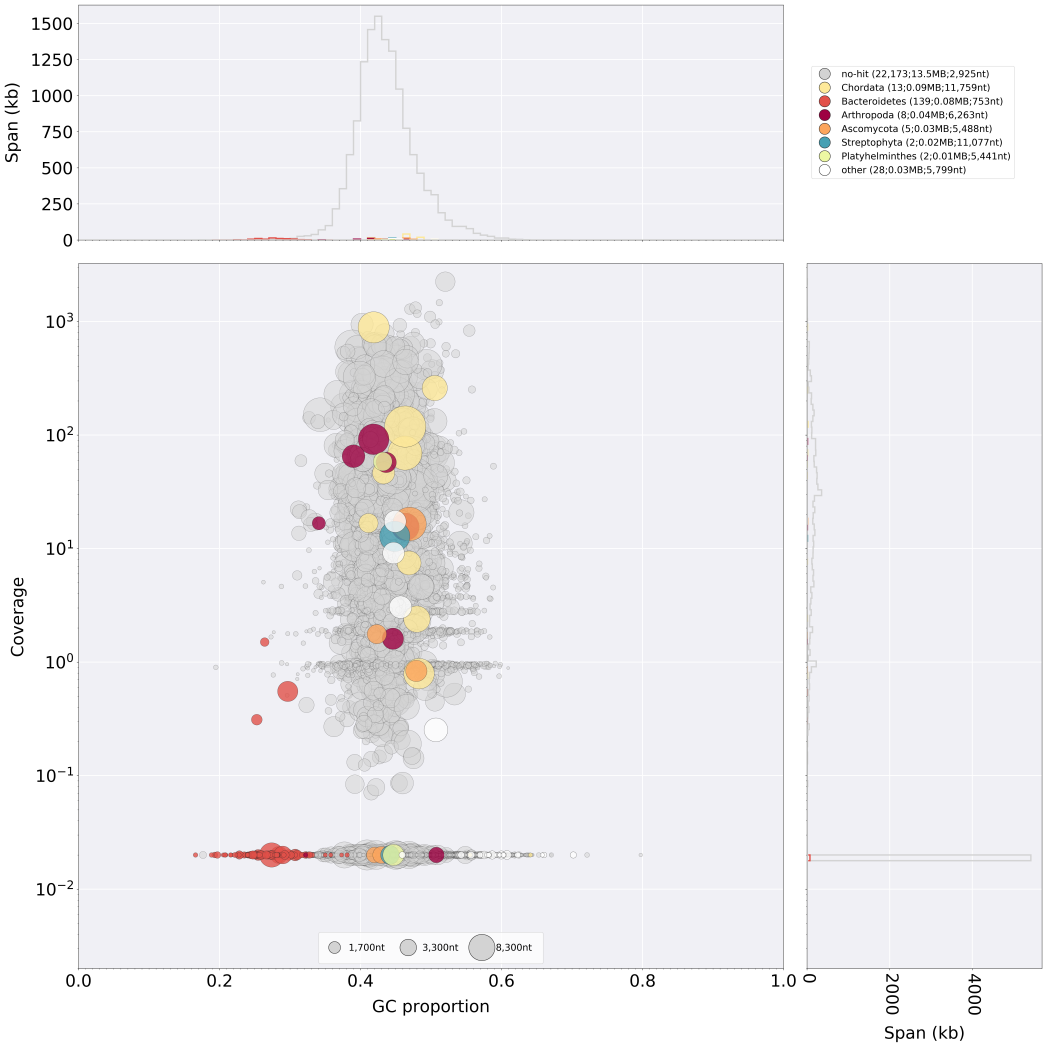
Read alignment properties

SciLifeLab

Contamination assessment

- Blobtools

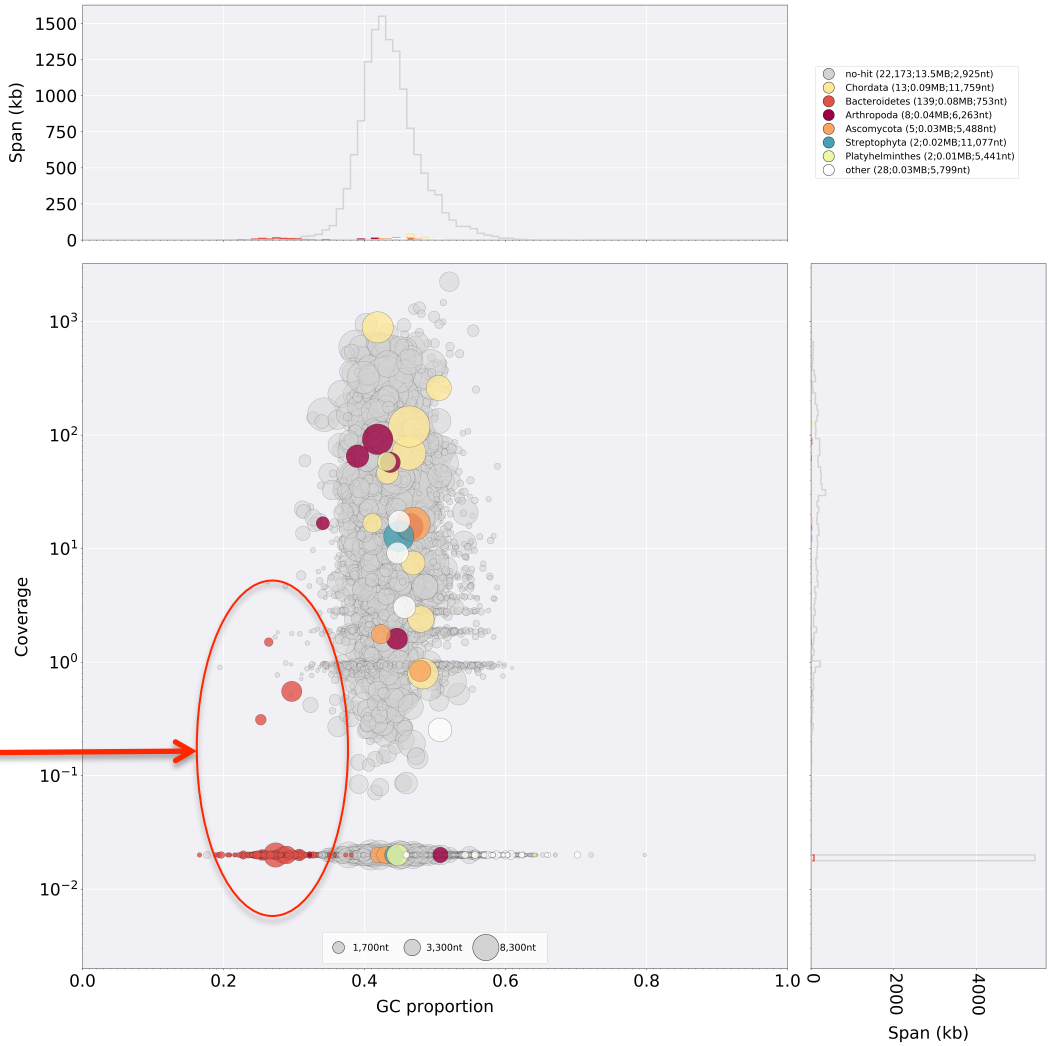
lidation/MiSeqSample1/MiSeqSample1_single_cell_spades_assembly_blob.MiSeqSample1_single_cell_spades_assembly_blob.blobDB.json.bestsu



Contamination assessment

- Blobtools

lidation/MiSeqSample1/MiSeqSample1_single_cell_spades_assembly_blob.MiSeqSample1_single_cell_spades_assembly_blob.blobDB.json.bestsu

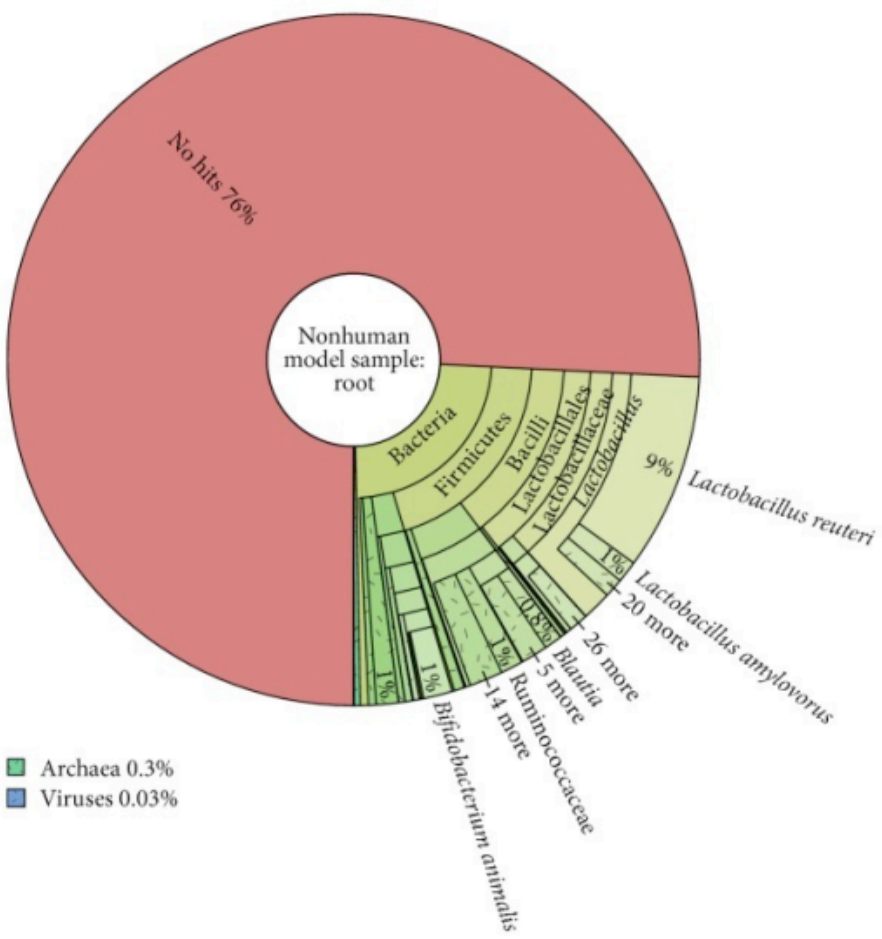


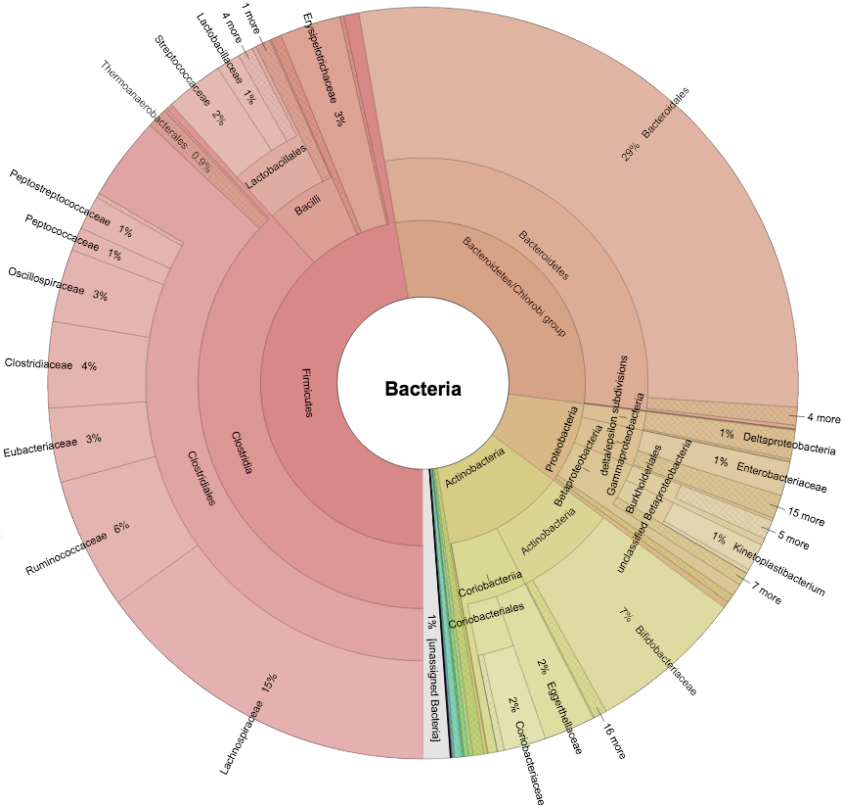
Sign of bacterial contamination

Contamination assessment

SciLifeLab

- Kraken taxonomic classification





Gene space statistics

SciLifeLab

- BUSCO *v3* provides quantitative measures for the assessment of genome assembly based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs.
  - Bacteria
  - Eukaryota
    - Protists
    - Metazoa
    - Fungi
    - Plants

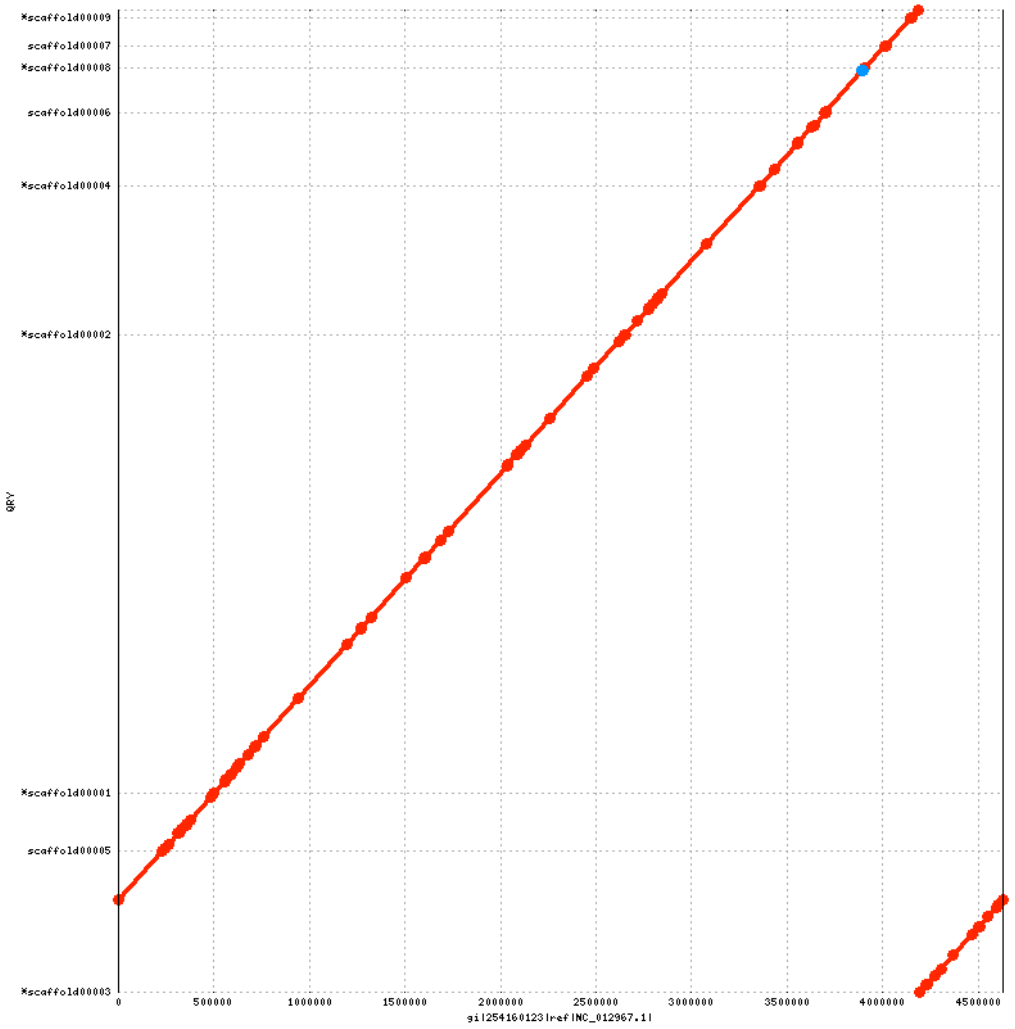
```
C:80.0%[S:80.0%,D:0.0%],F:0.0%,M:20.0%,n:10

8       Complete BUSCOs (C)
8       Complete and single-copy BUSCOs (S)
0       Complete and duplicated BUSCOs (D)
0       Fragmented BUSCOs (F)
2       Missing BUSCOs (M)
10      Total BUSCO groups searched
```

Comparative Alignment

SciLifeLab

- Dot plots (Nucmer, Gepard, etc)



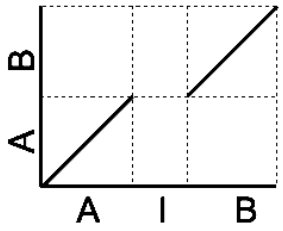
Comparative Alignment

SciLifeLab

- Dot plots

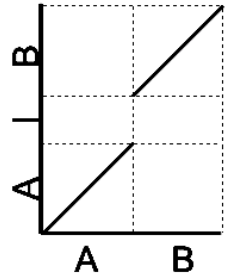
Insertion into Reference

R: AIB
Q: AB



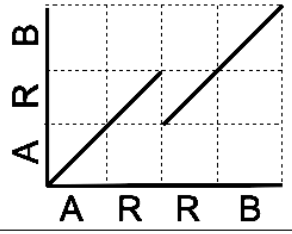
Insertion into Query

R: AB
Q: AIB



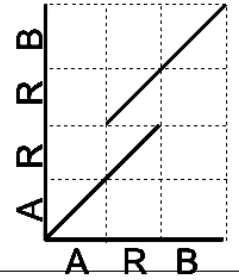
Collapse Query

R: ARRB
Q: ARB



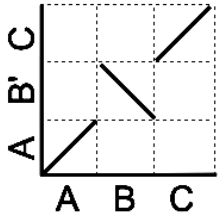
Collapse Reference

R: ARB
Q: ARRB



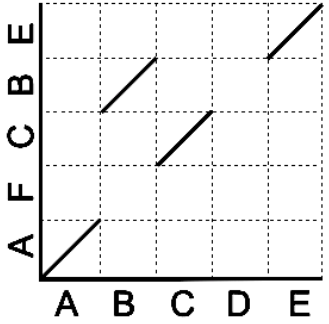
Inversion

R: ABC
Q: AB'C



Rearrangement
w/ Disagreement

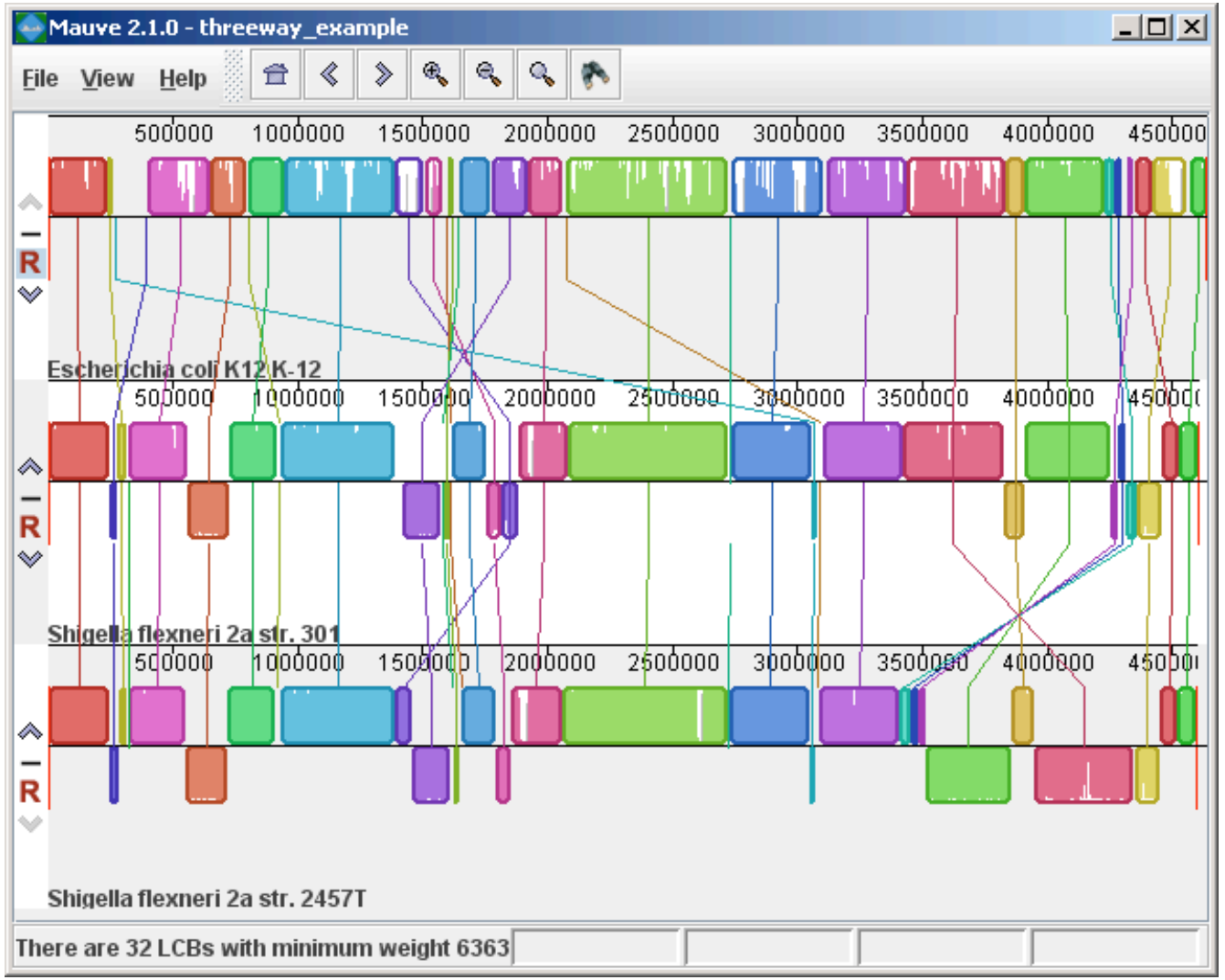
R: ABCDE
Q: AFCBE



Comparative Alignment

SciLifeLab

- Mauve



Improving your assembly

SciLifeLab

- Highly exploratory.

- Merging the best assemblies (assembly reconciliation) has been recently shown to not necessarily improve results.

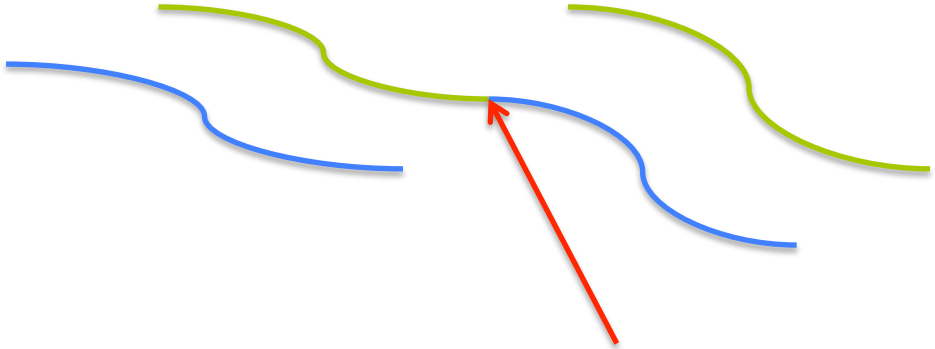
- Use the validation tools and assembly output to guide next steps.
  - Is coverage low after correction and trimming?
    - Change parameters to decrease minimum overlap depth
    - Increase overlap error rate
  - Is the assembly looking incomplete?
    - Can I get more data?
  - Is there contamination?
    - Align and filter

Correcting an assembly

SciLifeLab

- Downstream processing assumes correct assembly

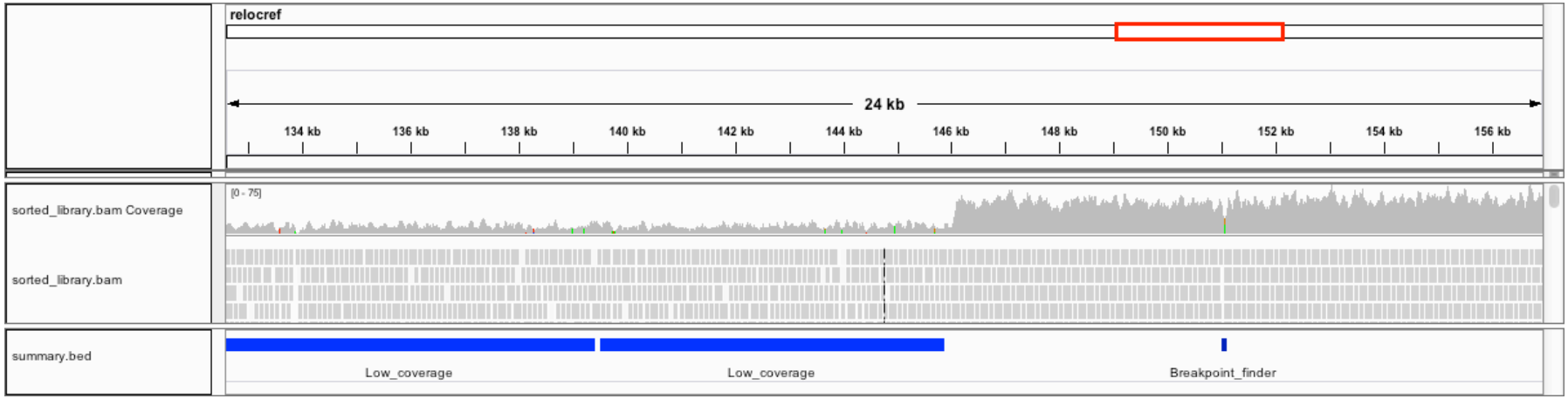
- Repeats and heterozygosity complicate assembly, however misassemblies are a primary reason for failing to improve assemblies further.



This misassembly prevents the contigs from being scaffolded correctly

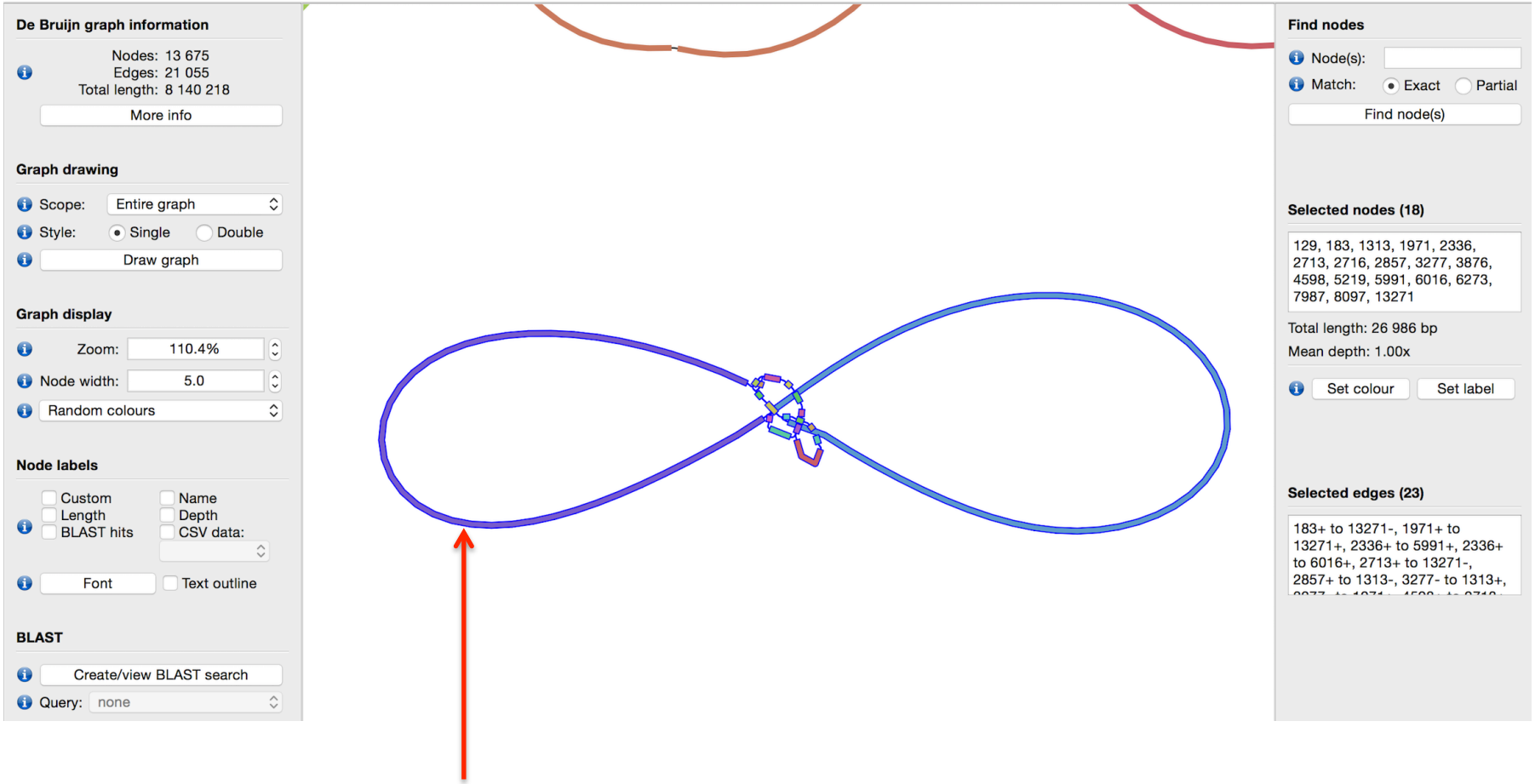
Correcting an assembly

- Manually breaking a contig
- (In a file editor)



- Programs
  - Reapr
  - [GAP5]
  - [QuickMerge]
  - [BIGMAC]

Data exploration



Evidence of COI gene found on this contig.
Hypothesis: This set of contigs make up the mitochondrial genome

Selecting the best assembly

SciLifeLab

- Illumina
  - Quast
  - Assemblathon_statistics
  - KAT
  - Bandage
  - Samtools flagstat
  - FRCBam / Reapr
  - IGV
  - Blobtools
  - Kraken
  - BUSCO
- PacBio / Nanopore
  - Quast
  - Assemblathon_statistics
  - Bandage
  - Samtools flagstat
  - (Bridge Mapper)
  - IGV
  - Blobtools
  - Kraken
  - BUSCO