

Genome Assembly Evaluation

Genome Biology (BIOL7263)
26Sept24

How can we evaluate the quality of our genome assemblies?

- Contig size and number
 - Quast
 - Bandage
- K-mer spectrum
 - Jellyfish and Genomescope
- Completeness
 - Quast - compare to known reference
 - BUSCO

Assembly statistics

Genome size - total # of bases in genome assembly

Contig/scaffold numbers - total number of assembled elements

ideally = # of chromosomes

Coverage - average read depth across the assembly

Assembly statistics

N50 - 50% of nucleotides in genome are contained in contigs that are greater than or equal to this length

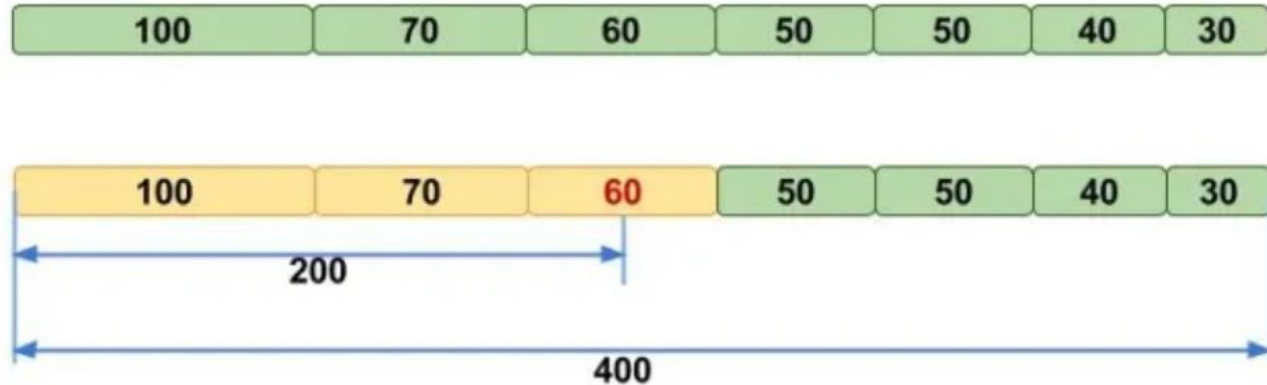


Fig. 1. Example of calculating N50 for a set of seven contigs. Here N50 equals 60 kbp.

Upper panel: Contigs, sorted according to their lengths.

Lower panel: Calculation of N50 using sorted contigs.

Assembly statistics

L50 - minimum number of contigs needed to contain 50% of the genome

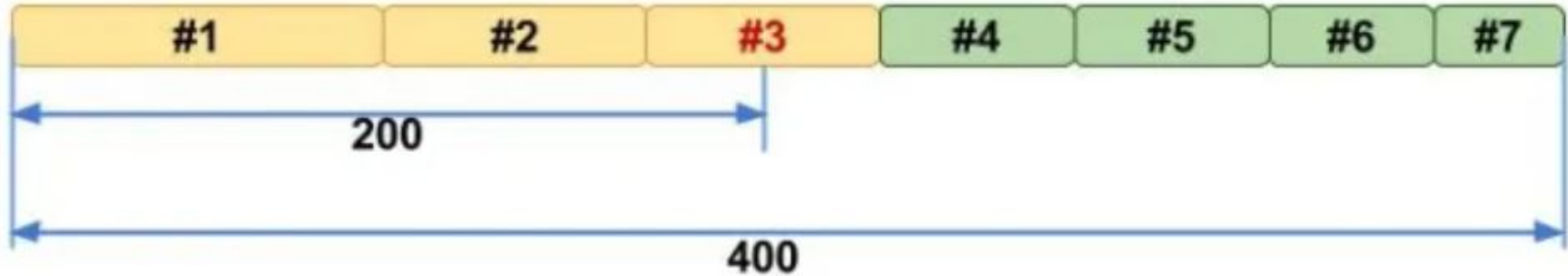
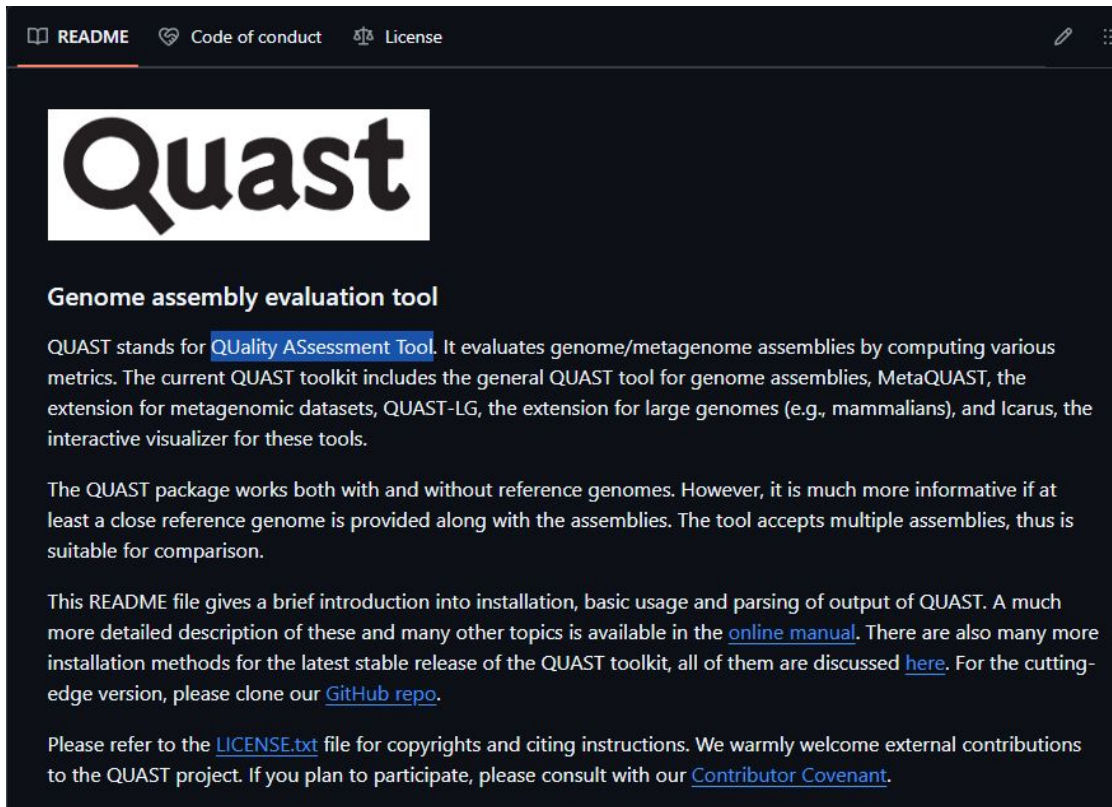


Fig. 3. Example of calculating L50 for the same set of contigs.

Here L50 equals 3.

Assembly statistics

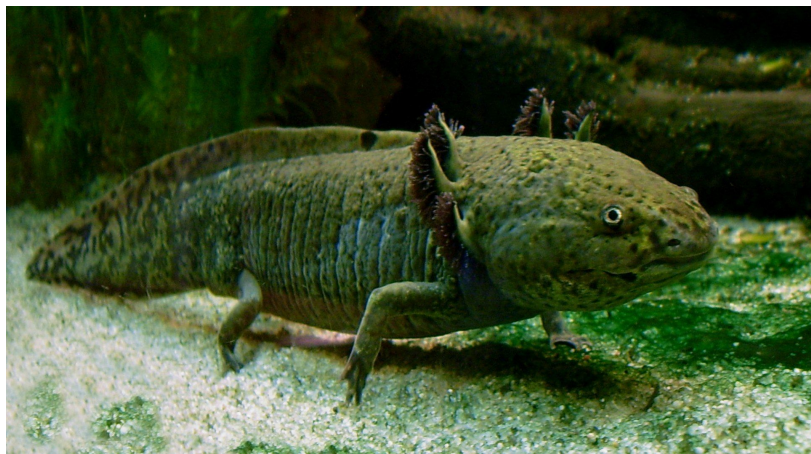


Walk through of quast
on our system:

https://github.com/mbtomey/genome_biology_FA24/blob/main/Lessons/Genome_Eval.md

<https://github.com/ablab/quast>

Assembly statistics



Ambystoma mexicanum

https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_040938575.1/

	GenBank
Genome size	29.1 Gb
Total ungapped length	29.1 Gb
Number of chromosomes	21
Number of scaffolds	220
Scaffold N50	1.5 Gb
Scaffold L50	10
Number of contigs	2,315
Contig N50	23.1 Mb
Contig L50	384
GC percent	46.5
Genome coverage	48.0x
Assembly level	Chromosome

Assembly statistics



Eurycea cirrigera

https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_035583035.1/

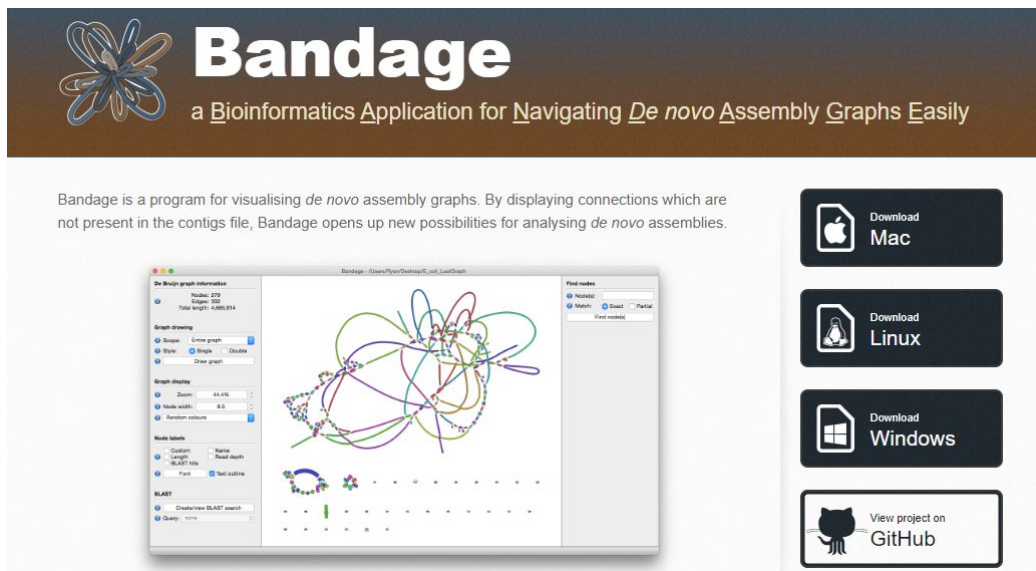
Genome size	689.9 Mb
Total ungapped length	689.5 Mb
Number of scaffolds	799,289
Scaffold N50	845 bp
Scaffold L50	249,826
Number of contigs	813,928
Contig N50	839 bp
Contig L50	252,580
GC percent	43
Genome coverage	60.0x
Assembly level	Scaffold

Visualize the assembly graph

- Understand the connections among your contigs
- Diagnose problematic regions of the assembly
- Visualize location of genes and other regions of interest

Let's review example here:

https://github.com/mbtoomey/genome_biology_FA24/blob/main/Lessons/Genome_Eval.md



Bandage
a Bioinformatics Application for Navigating *De novo* Assembly Graphs Easily

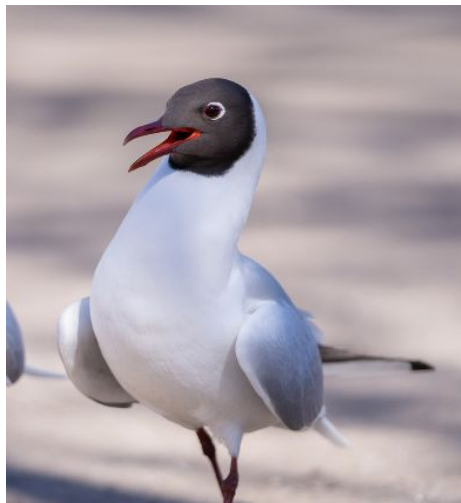
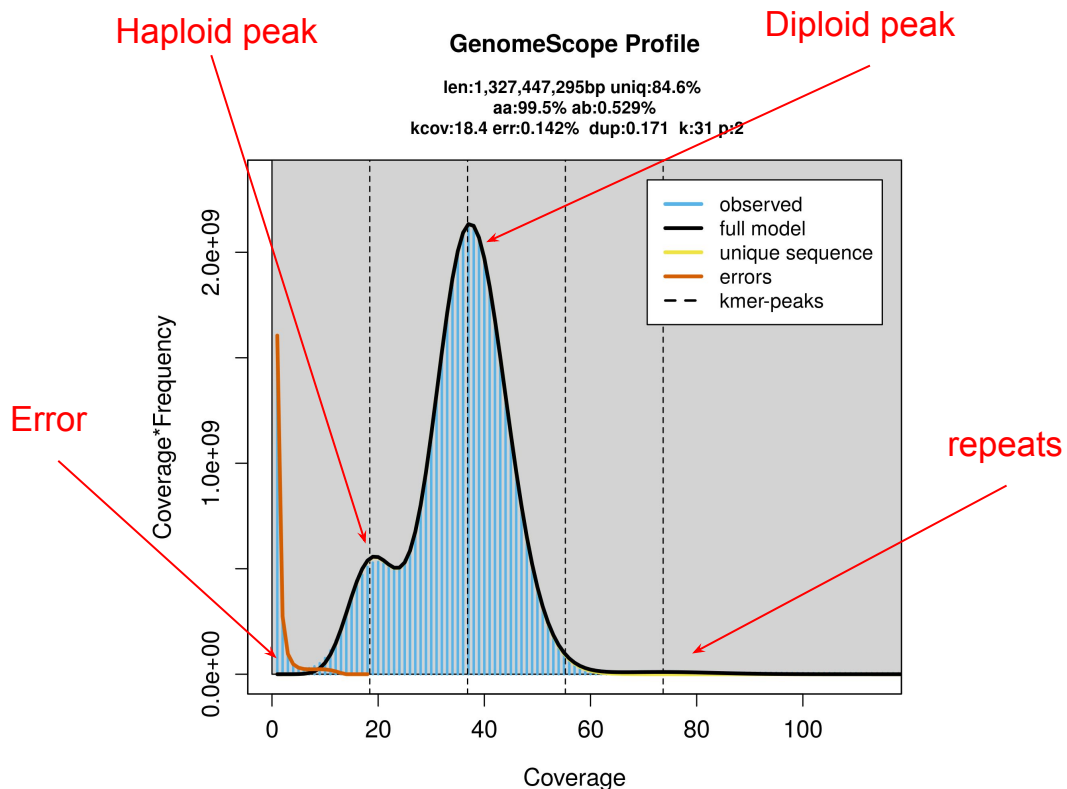
Bandage is a program for visualising *de novo* assembly graphs. By displaying connections which are not present in the contigs file, Bandage opens up new possibilities for analysing *de novo* assemblies.

The screenshot shows the Bandage application window. On the left, there are panels for 'De Bruijn graph information' (showing k-mer: 270, Nodes: 200, Total length: 4,889,814), 'Graph display' (with options for zoom, node width, and random colour), and 'Node labels' (with options for custom labels, name, read depth, and BLAST hits). The main area displays a complex assembly graph with nodes and edges. On the right, there are buttons for 'Download Mac', 'Download Linux', 'Download Windows', and 'View project on GitHub'.

K-mer spectrum analysis

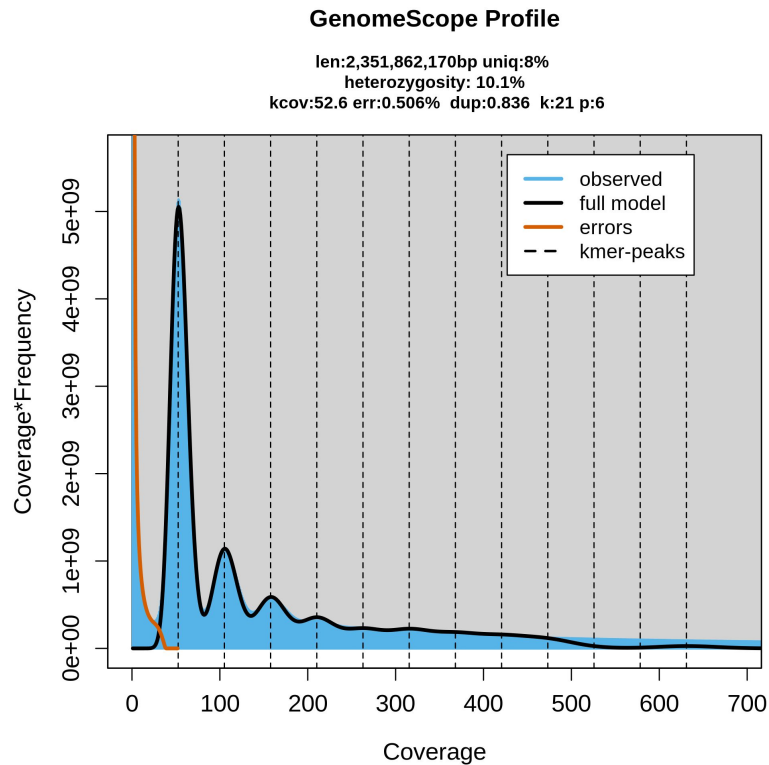
- Assess quality of your sequencing library
- Infer properties of the genome
 - Size
 - Ploidy
 - Heterozygosity
- Tools: Jellyfish and Genomescope

Typical vertebrate k-mer spectrum



- Diploid “p:2”
- Heterozygosity “ab:0.5%”
- Genome size - “len:1.3 Gb”

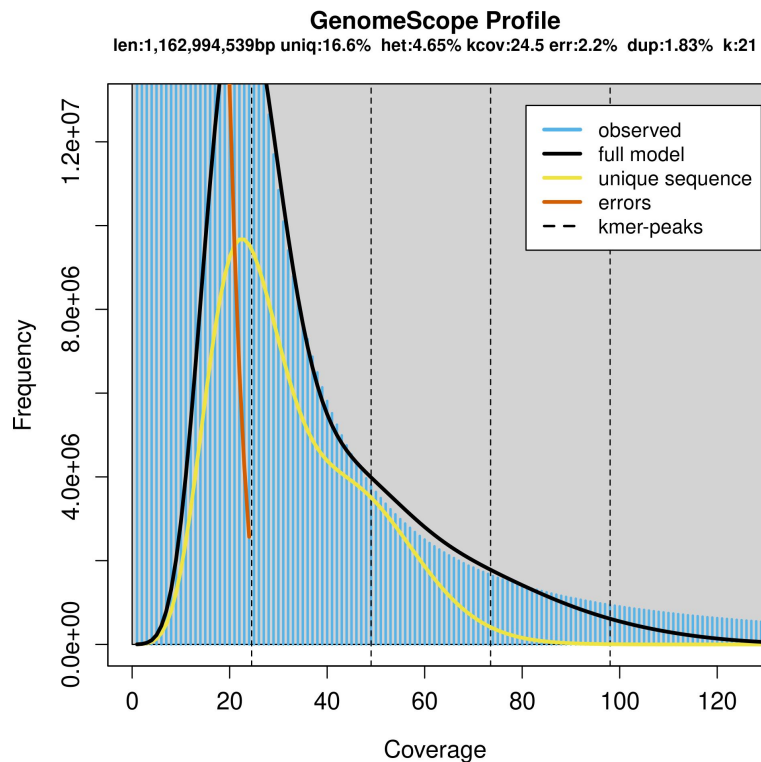
Spectrum of a hexaploid organism



Wheat (*Triticum aestivum*)



Spectrum with sequencing errors and/or contamination



- Large numbers of low coverage k-mers in library

K-mer spectrum analysis tools

- [Jellyfish](#) for K-mer counting
- Visualize with [Genomescope](#)

Let's try this with our own data:

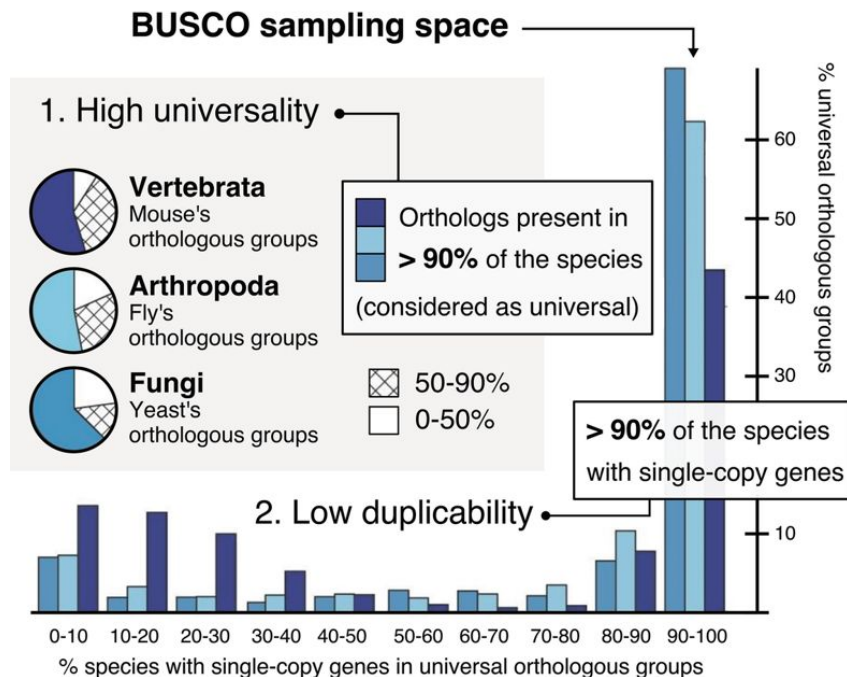
https://github.com/mbtoomey/genome_biology_FA24/blob/main/Lessons/Genome_Eval.md

Genome (transcriptome) completeness

- Quast - compare to know reference
- BUSCO

Benchmarking Universal Single-Copy Orthologs, BUSCO

- Ortholog - genes in different species that originated from a common ancestor and were separated by a speciation event



- Analysis relies on comparison of genome or transcriptome to a database of protein coding orthologs - <https://www.orthodb.org/>
- Available lineages - https://busco.ezlab.org/list_of_lineages.html
- Assumption - a complete genome or transcriptome should contain most of the BUSCO genes

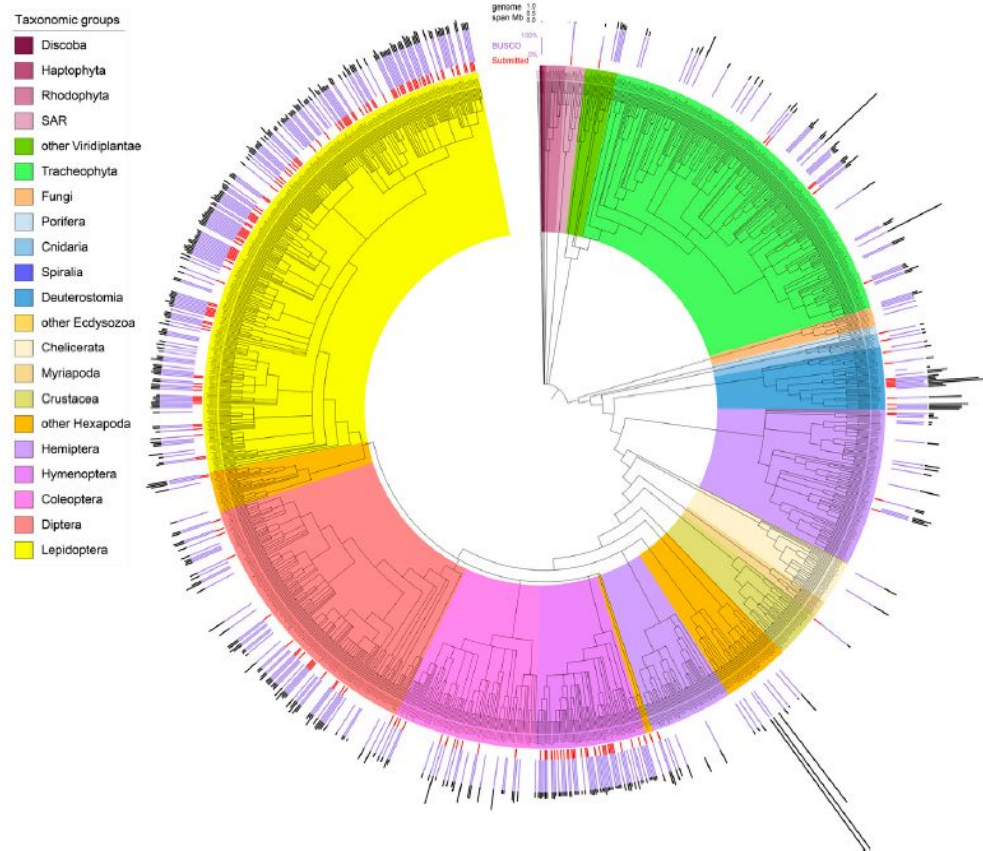
How BUSCO works

- Consensus sequences for each BUSCO are searched against the genome sequence using tBLASTn. Regions containing potential BUSCOs are identified. Up to three candidate genomic regions can be identified for each BUSCO.
- Candidate regions are extracted from the genome and [AUGUSTUS/Augustus](#) (or [Prodigal](#)) in combination with the BUSCO block profile is used for gene prediction. For transcriptomes, the protein prediction is used directly if available, otherwise the longest ORF within the transcript is used.
- Each predicted gene is then matched against the BUSCO group's HMM profile, sequences meeting the minimum alignment cut-off are considered orthologous.
- Orthologous sequences are then evaluated based on the expected-length cutoff. Sequences are classified as “Complete” if they meet the length cutoff, or “Fragmented” if too short. If multiple sequences meet the alignment and length cutoff they are classified as “Duplicated”. Any BUSCO without Complete, Fragmented, or Duplicated sequence is “Missing”.
- Finally, “Complete” sequences are used to build a new gene prediction model for Augustus (or Prodigal). A second round of Augustus (or prodigal) gene prediction is then performed on all BUSCO-matching candidate regions that did not yield a “Complete” ortholog. Classification is then carried out a second time on the new set of predicted genes.

BUSCO results

- **Complete** single-copy or duplicated match that has scored within the expected range of scores and within the expected range of length alignments to the BUSCO profile.
- **Fragmented** - matches that have scored within the range of scores but not within the range of length alignments to the BUSCO profile.
- **Missing** - no significant matches at all, or the BUSCO matches scored below the range of scores for the BUSCO profile.

What is a good BUSCO score?



A “good” assembly will usually have >95% complete BUSCO genes.

<https://www.pnas.org/doi/full/10.1073/pnas.2115642118>

BUSCO

Let's try this with our own data:

https://github.com/mbtoomey/genome_biology_FA24/blob/main/Lessons/Genome_Eval.md

BUSCO online: <https://gvolante.riken.jp/>