

Practical - Genome assembly QC

Here we will look at some genomes and assess their quality. Most of these genomes have (in my opinion) have quality issues. Remember that we are looking for as close to a perfect reconstruction of the original genome as possible, and we can assess this in different ways:

- Contiguity
- Correctness
- Completeness
- Contamination

These are explained here, [Quality control criteria for genome assemblies](#).

Exercise: Assessing genome assembly quality

The exercise is to download some assembled draft genomes and assess if they pass routine quality control. *The genomes are available at <https://zenodo.org/records/10018484>*. You can directly download them on the command line with `wget` or similar.

```
wget -O additional_genomes_for_qc.zip
https://zenodo.org/records/10018484/files/additional_genomes_for_qc.zip?
download=1
```

The files are FASTA (they are plain text) inside the zip file. There are eight genomes. Each FASTA file will look something like this. There will be multiple contigs, one after the other. Each with a fasta header (`>NODE_XXXX`) followed by the nucleotide sequence (`ATGC`).

```
>NODE_126_length_2252_cov_16.6409_ID_251
CAGCGTGGACTGATGTTTCAG.....
>NODE_22_length_135487_cov_12.0245_ID_43
GGCCGAGGCTCCCCACCGCGCGGG...
>NODE_68_length_16957_cov_12.5198_ID_135
TGGTGTGGTGCCAACGGCCTGACC...
>NODE_16_length_182200_cov_11.9821_ID_31
GCCGCTTTTCGCGTTGCTTAATCT...
```

Try to create a table of your assessment (with better explanations) like the one below:

Sample name	Pass/Fail	Reason
sample_1	Yes	
sample_2	No	
sample_3	Maybe	
sample_4	I don't know	
sample_5	Could you repeat the question?	
sample_6		
sample_7		
sample_8		

This exercise is open-ended, and you can use any tools you like. You can use the tools mention in [Quality control criteria for genome assemblies](#), or you can try other tools. You can work in groups, or divide the different criteria between yourselves.

[Here is a link to the assemblies](#)

If you do not have time, or the capacity to run these analyses, please use the precomputed results below. It is more important that you learn how to interpret the results.

- [BUSCO - summary images](#)
- [BUSCO - short summaries](#)
- [Kraken2 - Reports](#)
- [MLST - summary](#)
- [QUAST - pdf](#)
- [QUAST - table](#)
- [QUAST - html](#)

If you complete the task above, also try to answer the following questions:

What is N50?

What is GC Content?

What is "Genome Coverage"?

How does BUSCO measure completeness?

How does aligning to a reference genome help assess completeness?

How does using Kraken help assess contamination?

[Answers to exercises](#)