Code to genomes

Amy H Fitzpatrick

2023 Feb 08

Quality control

Quality Control

NanoPack

Quality assessment of raw reads

Filtlong

Filter low quality reads with low quality scores

Remove Host reads

Use a package such as **Hostile** to remove reads aligning to the host

Genome assembly

- Canu
- viralFlye

Genome annotation

- VAPid
- Prokka

Taxonomy

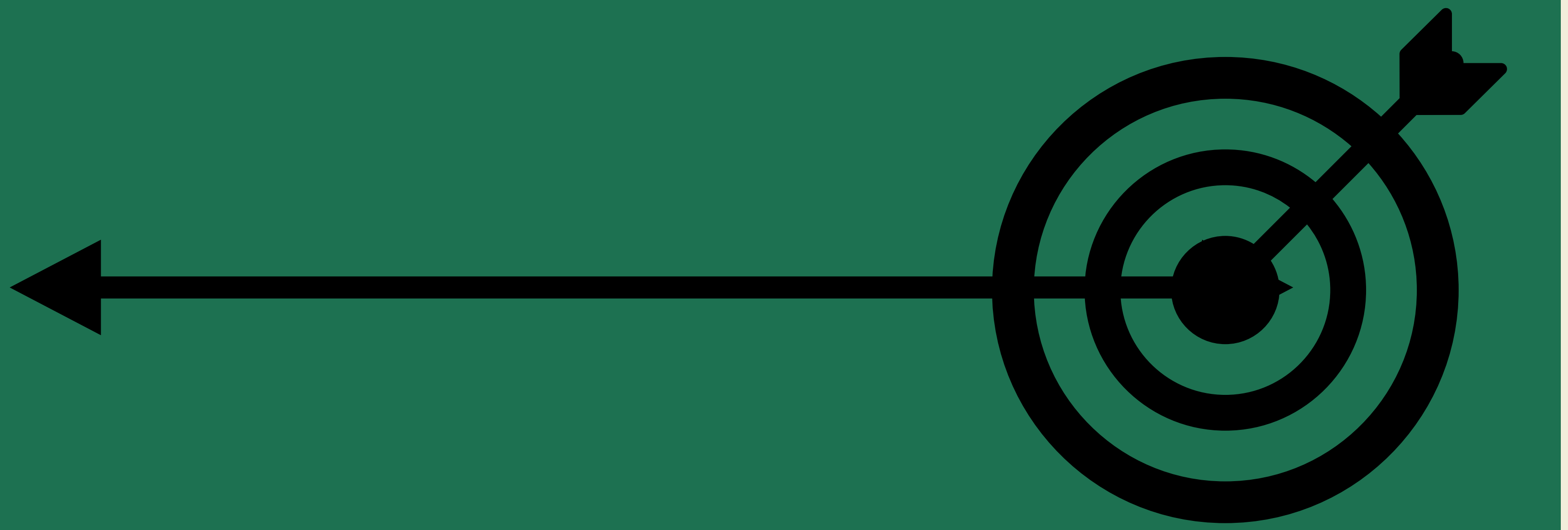
- Kraken2
- viralVerify

Removal of Short or Low-Quality Reads

- Low-quality reads contain errors introduced during sequencing (e.g., base-calling errors).
- Low-quality reads introduce errors like false overlaps or incorrect alignments.
- Low-quality reads lead to misassemblies, chimeric contigs, or incomplete assemblies.

Removal of Short or Low-Quality Reads

- Filtering based on quality scores (e.g., Phred scores) removes error-prone reads, improving downstream accuracy.
- Filtering before assembly enhances accuracy and completeness of assembled sequences.
- Removing low-quality reads reduces assembly errors, ensuring reliability.
- Filtering by size allows selection of desired DNA fragment sizes (e.g., for targeted sequencing).
- Focuses analysis on relevant genomic regions.



How will you filter your reads?

Task 1

Before writing your script with filtlong, you need to define your filtering metrics.

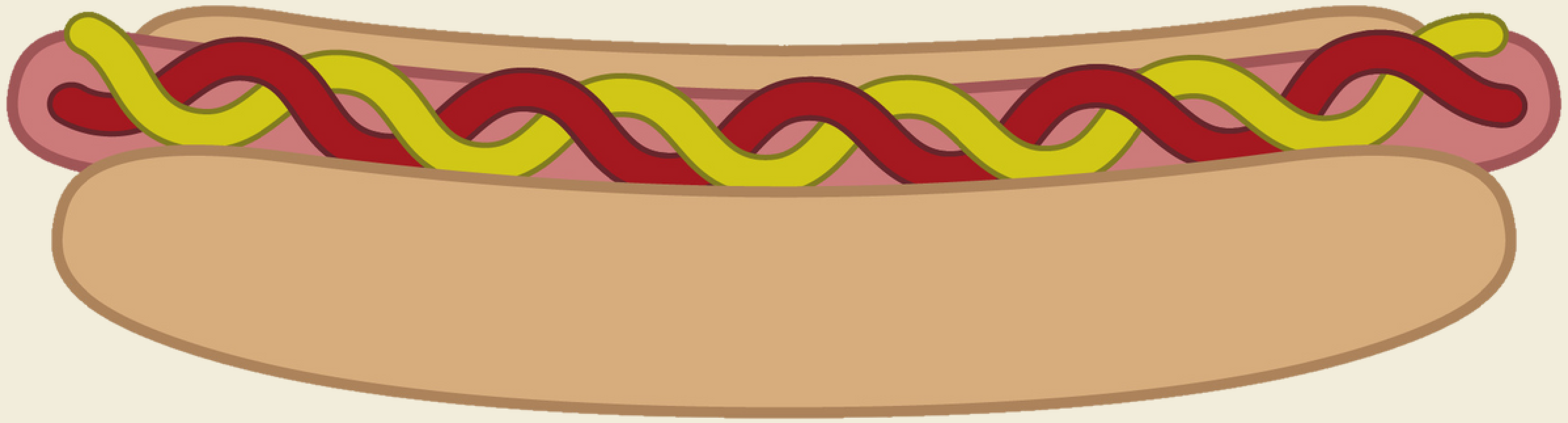
- Where do the majority of your reads lie in terms of length and quality?
- What was the goal (see publication) for sequence length? What questions are they trying answer?
    - How would non full length viral sequences impact their ability to answer the question?
    - Does length matter?
- What kits from ONT were used to generate the data? What are the associated quality inplications?

Install and run filtlong

Task 2:

Now that you have defined how you will filter your data, start drafting a SLURM script to filter your data with the tool Filtlong.

- You will need to install the tool using cmake compiler, see the github page for instructions
- Create a new slurm script and send the filtered results to a new folder
- Run NanoQC on the output and compare it to the previous data



Host genome removal

Task 3:

Review the publication for information on the potential host genome. Is there a reference genome available for that host?

- Check if there is a clear host?
- If so, find the GenBank accession code for the reference genome
- Practice what you learnt in week one and download the reference genome using entrez-direct
- Install a new conda environment Hostile