Data Science in Bioinformatics

Day 8

This is going to be the last exercise sheet, so we are going to work on exercises concerning the last day of student presentations. In the presentations the topics of assembly control and lineage assignment were addressed. These would also be the final steps of the pipeline you are going to create as the final project in this module.

1) **Assembly Control**

    a. What is the input needed for the assembly control with QUAST?
    b. Run QUAST based on the assembly you created in the last exercise with MEGAHIT and use the Coronavirus reference genome Wuhan as reference genome.
    c. Hand in the report.html file QUAST is creating as output.
    d. What is the QUAST report telling you how big the largest alignment is?

2) **Variant calling and lineage assignment**

    a. Why is it important to call lineages for, e.g. samples of the corona virus?
    b. What are SNVs, SNPs and Indels?
    c. What input is needed for variant calling?
    d. Use your alignment from day 6 and perform a variant calling with samtools/bcftools. What input is needed? Which preprocessing step has to be done before calling variants?
    e. Use your scaffolded sequence from exercise 7 and perform lineage assignment with pangolin via the command line. What does the output look like and which lineage was found?