Data Science in Bioinformatics

Day 7

This exercise sheet is going to address the topics of the last presentations, assembly, and scaffolding. Both are essentially important if we want to be able to call the kind of variants we find in a sample.

1) **Assembly**
    - **a.** What does assembly do?
    - **b.** What is the difference between de-novo and reference-guided assembly?
    - **c.** What is the algorithm that is used by MEGAHIT? Describe this algorithm in your own words.
    - **d.** What is the result of a genomic assembly and what is the difference between it and a read?
    - **e.** Run MEGAHIT with the samples you created during filtering in the last exercise.
    - **f.** How many contigs do you get in the final assembled output fasta file?
    - **g.** What do the header parts “flag” and “multi” mean?

2) **Scaffolding**
    - **a.** When is scaffolding needed?
    - **b.** Scaffolding with RagTag: which are the four modes the tool can be used with?
    - **c.** Choose the best option to process the contigs from e), what is your result?