F2022 Long Assignment — Genomic project. A heterogeneous bacterial mixture

Input files are in **/media/Data_1/F22_BIOL550_LA/** on Mozart; these files are too large to put
on Blackboard

You received a genomic dataset from collaborators consisting of Illumina (paired ends) and Oxford Nanopore sequencing data. Your role in the project is to assemble and annotate the genome. Your collaborators are expecting data from a *Staphylococcus aureus* strain (**expected genome size ~ 2.8 Mbp**), but hey, who knows what is in there!

Automation with Bash and/or Perl scripts could save you a lot of time…

10 pts. *Read processing*

i. Investigate the read quality of the Illumina and Oxford Nanopore datasets.
The error below can show up with FASTQ files containing long reads:
Exception in thread "Thread-3" java.lang.OutOfMemoryError: Java heap space
If so, use more threads with -t, *e.g.* -t 4

ii. Plot the read length distribution / calculate the metrics of the nanopore dataset.
You can use my python script to do that:
https://github.com/PombertLab/Misc/blob/main/read_len_plot.py
Read lengths are very important for 3rd generation platforms. Because read sizes
influence the assemblies, we should always verify their overall lengths

iii. If required (and possible), filter accordingly.
When using nanofilt, you can safely ignore the warning about pandas

20 pts. *Genome assembly*

iv. Assemble each dataset using SPAdes (Illumina), Shasta (Nanopore; read the manual; does not work with .gz files), Flye (Nanopore) and Unicycler (Illumina + Nanopore; read the manual). Remember that different parameters may lead to different results. **These processes can take a long time**, make sure to use **screen** (screen -S name_of_screen) to run the analyses so that you can detach (ctrl+A+D), then re-attach (screen -r name_of_screen) to the shell as needed.

20 pts. *Contamination removal (from assemblies)*

v. For each assembly, identify the staphylococcal contigs using BLAST homology searches against the ref_prok_rep_genomes database (located in /media/Data_1/NCBI/REPGENOMES), then parse out the results with Perl to keep only the contigs that are from *Staphylococcus aureus* (see runTaxonomizedBLAST.pl/ parseTaxonomizedBLAST.pl from **BIOL550_06d_Assemblies.pdf**'s Exercise 11). These scripts are available on Mozart for you to use.
To save computation time, use the megablast algorithm rather than the default one
(blastn). Because a representative genome for this staphylococcal species might not be in
this database, assume that any *Staphylococcus* hit is a valid one.

20 pts. *Assembly comparisons*

vi. Compare the various assemblies (before and after filtering) using Quast. Select the best assembly for downstream analysis. Explain why you selected this assembly. Does it match the expected size? Based on your knowledge, also explain which sequencing approach appears better for this project and why.

15 pts. *Genome annotation*

vii. Use PROKKA and DFAST to annotate the best *Staphylococcus aureus* assembly (without the contaminants). Use F22 as the locus_tag prefix and autoincrement values by 10. For PROKKA, enforce GenBank compliance.

viii. Compare the total number of proteins predicted by PROKKA and DFAST for your assembly. Are the results congruent? Explain your answer.

15 pts. *Contamination removal (from sequencing datasets)*

ix. Let's filter out the sequencing datasets to keep only the reads from *Staphylococcus aureus*. Use your best assembly as reference. We can use get_SNPs.pl (--rmo + --bam), minimap2 and samtools bam2fq for this purpose. GZIP your FASTQ outputs. Compare the number of reads before/after with FASTQC. Compare the *.fastq.gz sizes with du -sh too.

Important:

- You can perform this assignment either **alone or as a team of up to 4 people**. The assignment is perfectly doable by yourself, but people may want to team up with students from other discipline to facilitate interpretation of the data.
- **Each student can use up to 4 threads** for computations: team of 4 => 4*4 = 16 threads total
- **If you want to perform this assignment as a team**, contact us by email with the names of your team members (CC all team members) and we will create a group and a shared directory for your team on Mozart to facilitate teamwork.
- **Some of these computations will take hours to run**; do not wait at the last minute to perform this assignment.
- Remember to use screen for long computations so that you can detach/reattach and disconnect from the server while running the analyses.
- **You can ask questions about the assignment**.

Submit on Blackboard:

- A **report (in PDF format)** describing the steps and command lines you did, with visual support whenever useful (*e.g.* screenshots/images of read length distributions and FASTQC/QUAST reports are informative and thus expected). Using a font like `consolas` combined with a different color scheme can be very useful to show command lines. Also include in the report brief descriptions of issues you encountered (if any) such as bugs.
- If submitting as a team, include all team members on the front page of the report.
- If you created Perl and/or Bash scripts, compressed them as a GZIPed tar archive and submit it together with your report.

Deadline: Dec 9th, 2022.