Analysis log

20240709

Received task. Today and tomorrow will do the planning, but also will prepare a data analysis plan. Will try to drop in for the HPC garage to ask if I can use University of Helsinki resources for the alignment.

20240712

On Wed, 20240710 talked to HPC team, made sure I can use the cluster. Also, assembled in my head approximately how I want to realise the alignment and QC parts.

- The base pipeline is going to be: FastQC > MultiQC > Trimmomatic > FastQC > STAR > Cufflinks ? > bowtie (PhIX contamination) > bwa and samtools for rRNA contamination > kraken for bacterial contamination
  - Check trimmomatic manual — type of adapters?
  - use fastp?
  - Will I do quality trimming?
  - Do I need to install parallel?
  - Check e-mails from Biomics - contamination? trimming reads? pipeline?
  - Check MHH facility files prepared for Sasha - pipelines? etc
  - Check paper BG cite about the pipeline
  - Check the BG phenotyping paper
  - Check TCGA guidelines
  - Check GENECODE guidelines (or what?)
- When I'm sure about the tools, write bash scripts for all the tools
  - Cluster management and environment:
    - Will it be run directly on a cluster?
    - Will I make a script for environment creation or just share env yml?
    - Will I use snakemake?
    - Easy replacement of tools!!
- Data access scripts:
  - Will I call them from python notebook?
  - Just plug in the dataset GSE and it will download all the FASTQ files
  - Prob just launching a bash script via subprocess.Popen()