**Getting from sequencing reads to a table of read counts**

**Typical workflow:**

* Do a quality control check of the reads data to see if any trimming is required
* Generate an indexed version of the human reference genome sequence
* Align the RNA-seq reads to the genome, outputting BAM files
* Generate a counts table displaying the numbers of reads that represent each gene

The workflow for performing either variant analysis or differential gene expression analysis starts with obtaining the fastq files. When you extracted them from the GEO database, the only quality check you did was to make sure that paired-end read files were matched in size. Now you want to look at the overall quality of the reads themselves.

1. Write a script to generate quality control readouts with the program fastqc. Interpretation of the readouts is the topic of a different worksheet, so we'll ignore this step for now.

2. What is the path for your indexed human genome files?

path:

If you do not have the indexed files already, write the UNIX command for copying them from the course files to your own directory of choice:

command:

3. Now we need to map the reads to the genome. We've used two different programs for this purpose – bowtie2 and STAR. Which is the one to use when working with organisms that have introns?

4. Write an SBATCH script to map the reads to the genome. This will generate a bunch of bam files (compressed versions of sam files).

5. Write a script to generate a counts table. The command you'll use is part of a program called **subread**, which is available as a conda environment on our course site. Be sure to include the appropriate lines in your script for activating the subread environment. Also, your bam reads should all be in a directory called "bams", at the same level as your genome annotation file. Here's the subread command for generating a counts table:

featureCounts -a Homo\_sapiens.GRCh38.115.gtf -p --countReadPairs -o counts.txt -T 8 bams/\*.bam

Output will include these files in the directory you specified in the command above.

A black background with green letters

Description automatically generated

Use the OOD interface to download the table and open it in Excel. You can delete the columns headed chr, start, end, strand, and length, leaving only the columns of counts for each of the samples. If there is a row above the sample names, you can delete that, too.

Save the spreadsheet as a .csv file. (You have now recreated the counts\_data.csv file that you read into RStudio to create a PCA plot, perform DGE analysis, and generate a heatmap. You now know how to do DGE analysis on RNA-seq data, start to finish!).