**Mustafa Cerit Homework\_05\_RNAseq\_analysis**

ceritmus@gateway-03:~$ mkdir 05\_RNAseq\_01

ceritmus@gateway-03:~$ cd 05\_RNAseq\_01

ceritmus@gateway-03:~/05\_RNAseq\_01$ wget <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR357/053/SRR35753274/SRR35753274_1.fastq.gz>

and other also dowloanded

ceritmus@dev-amd24:~$ module load Salmon/1.10.1-GCC-12.3.0

ceritmus@dev-amd24:~$ salmon index -t Zea\_mays.Zm-B73-REFERENCE-NAM-5.0.cdna.all.fa.gz -i Zea\_mays\_Salmon\_index -p 8

Version Info: This is the most recent version of salmon.

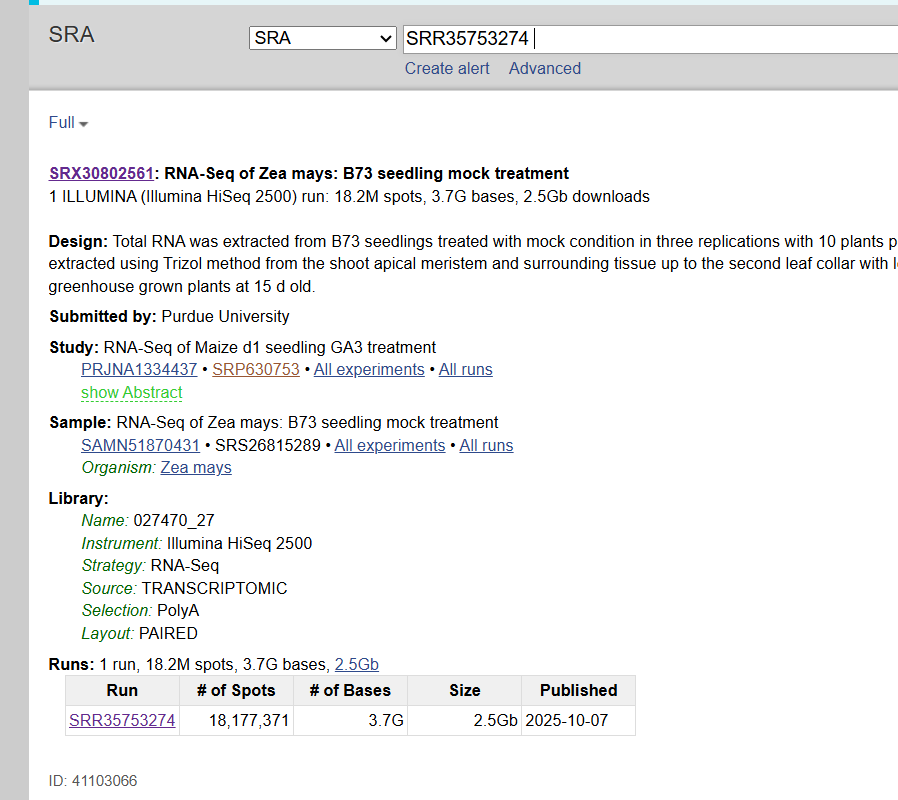
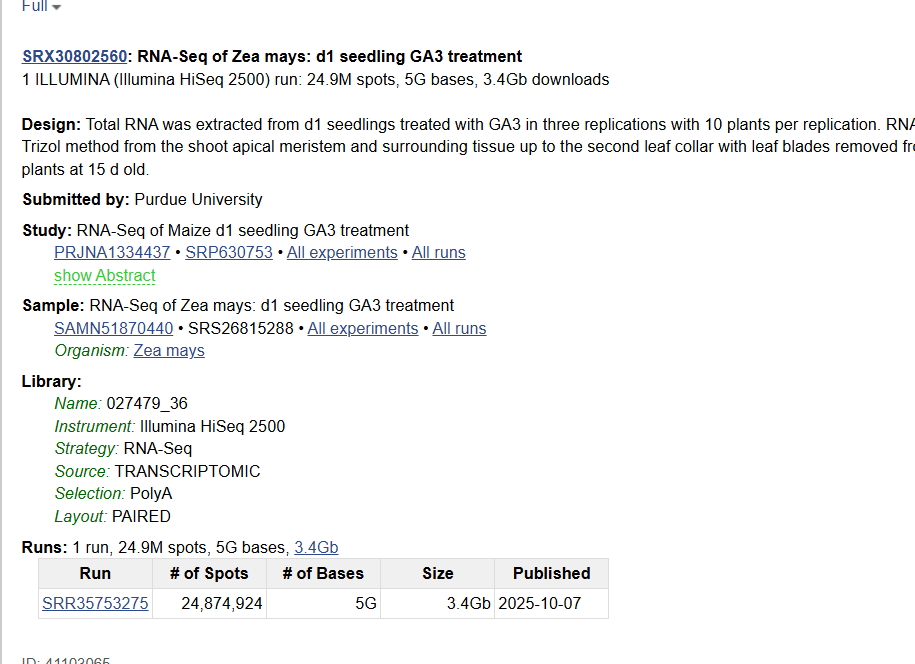
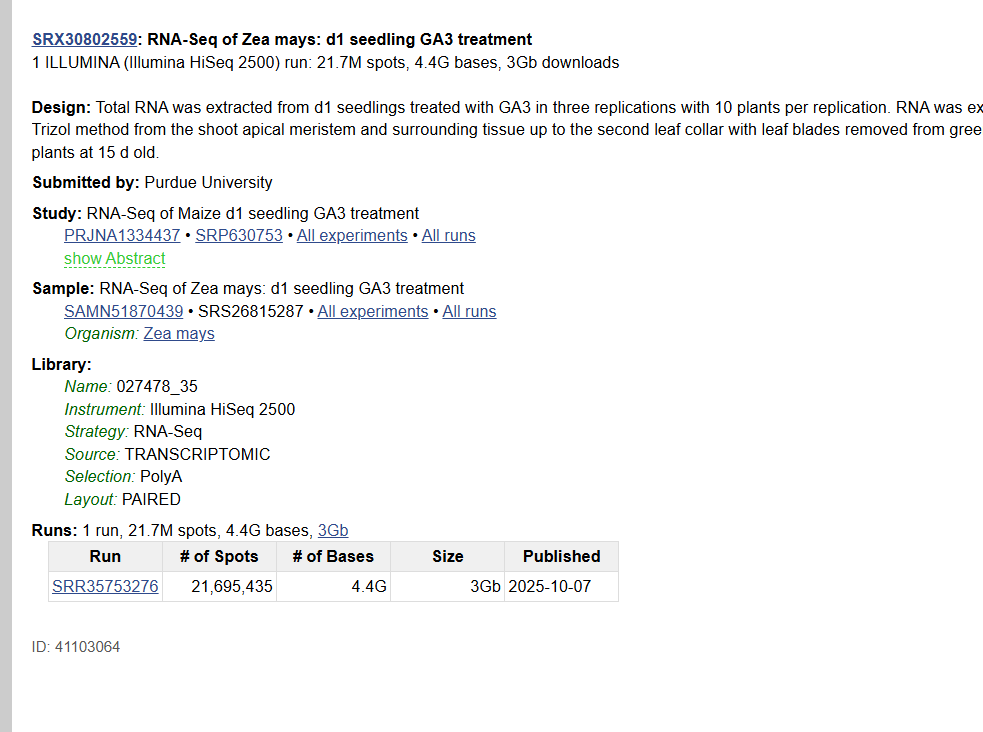
Activity

Let's run salmon quant on some of the maize RNAseq samples under /mnt/research/PLB\_812\_F25\_001/05\_RNAseq\_I/RNAseq\_reads. There are five RNAseq samples here, please run salmon on three of them and answer the following questions below.

1. Look up the samples on the Sequence Read Archive (SRA) from NCBI. Use the file names (e.g., SRRXXXXXX) to search for the samples in the NCBI SRA: <https://www.ncbi.nlm.nih.gov/sra>

1a. What tissues are these three samples derived from?

Answer1a. These 3 samples are derived from maize seedling tissues.  
They include one mock-treated sample and two GA3-treated seedling samples.

1b. Can you determine the experimental design (e.g., biological replicates, time points, or treatments) based on the metadata provided on NCBI?

Answer 1b. The experimental design includes one control (mock-treated) and two GA3-treated maize seedling samples, representing biological replicates sequenced using Illumina HiSeq 2500 (paired-end RNA-Seq).

1. Build an index for maize using the code provided above. How many transcripts are included in your index? (Hint: You can find this information in the log file generated during index creation, or by using tools like grep to count entries in the FASTA file.)

Answer 2. Approximately 17,192 unique transcripts were indexed after removing duplicate sequences.

1. After running Salmon on three samples, compare the read mapping statistics for each sample. You can find this information in the logs subdirectory of each output folder (e.g., SRR940276.quant/logs/).<br) 3a. Which sample had the highest mapping rate?

Answer 3b. The sample SRR35753274 (mock treatment) had the highest mapping rate of about 90.98%, indicating the best alignment quality among the three samples. The other samples showed mapping rates of 87.29% (SRR35753275) and 87.99% (SRR35753276).

3b. Do you think this mapping rate is acceptable? (Typically, a good mapping rate for RNA-seq data is above 70%. If your rate is lower, consider potential reasons.)

Answer 3b. Yes, the mapping rates are acceptable. All three samples have mapping rates more than 85%, which is well above the typical threshold of 70%.

3c. What could the reads that don’t map to the genome or transcriptome represent? (Consider possibilities such as sequencing errors, contamination, unannotated genes, or non-coding RNA.)

Answer 3c. The unmapped reads might be caused by sequencing mistakes, contamination, or regions of the genome that are not included in the reference. They could also come from unknown or non-coding parts of the RNA.

1. Open one of the quant.sf files either in the terminal (using less or cat) or in a text editor. Focus on the Name (gene ID), TPM (transcripts per million), and NumReads (number of reads) columns.

4a. What is the range of TPM values you observe?

Answer 4a. The TPM values range from 0 to around 3200

4b. Are there many entries with 0 TPM? If so, why might this be the case?

Answer 4b. ceritmus@dev-amd24:~/05\_RNAseq\_01$ awk '$4 == 0 {count++} END {print count}' SRR35753274.quant/quant.sf

24813

This happens because some genes are not expressed in this sample. Other reasons could include low expression

4c. Are the TPM and NumReads values whole numbers? Why or why not?

Answer 4c. TPM and NumReads are not whole numbers because they are calculated and normalized, not just counted directly. it makes easier to compare gene expression between different genes and samples.