Homework/Mini Project 3

W. Evan Johnson, Ph.D.
Director, Center for Data Science
Rutgers New Jersey Medical School

Due May 30, 2023

Now its time to practice what we have learned in class and learn even more! For this homework/mini project you will do a RNA-seq analysis of the TB/HIV dataset. Note that your homework should be written in R Markdown, and turned in by uploading a tarball with your .Rmd, .html (from Rmarkdown, MultiQC, etc) and other outputs on Canvas.

RNA-sequencing analysis

1. To access the data for Homework 3, you will have to download the data from the Sequence Read Archive (SRA) using the `sratoolkit`. There are 33 fastq files, with the SRR numbers are listed in the `homework3_srr.txt` file.
2. Align the reads to the human genome reference using your choice of the `Rsubread` or the `STAR` aligners.
3. Use the `featureCounts` function in the `Rsubread` package to generate a counts file for this dataset.
4. Generate a `SummarizedExperiment` object for your counts. The `colData` for these data are provided in the `homework3_metadata.txt` file.

5. Preprocess these data by removing `TB-HIV-ART` samples (should be two of them), removing any genes with 0 expression for all samples, and by generating a log counts per million assay.
6. (Extra credit) Create a batch corrected assay in your `SummarizedExperiment` using ComBat-Seq. You can use the `ComBat_Seq` function in the `sva` package, or simply do it in `BatchQC` and then extract the `SummarizedExperiment`. For this example, pretend that `disease_status` is the batch variable. Note that this will remove the disease status variability, so **don't** use this assay in the following analyses! This was merely a practice for cases where you have an actual batch variable.

7. Apply SVA and UMAP to your data and generate dimension reduction plots for the results. Color the TB-HIV to the HIV only samples in different colors. Note that you should be using the log CPM values for this analysis.
8. Use `DESeq2` to do a differential expression analysis (on the counts) comparing the TB-HIV to the HIV only samples. Provide the top 50 most differentially expressed genes.

9. Now conduct the same analysis using `limma` on the log CPM values. How do the `DESeq2` results compare to the `limma` results?
10. Give a heatmap plot of either the `DESeq2` or the `limma` results (top 50). Add a colorbar for disease status.
11. Conduct a pathway analysis of the top 50 genes usign a tools such as enrichR (through R or online). What are the top scoring pathways?
12. (More extra credit) Conduct a TBSignatureProfiler analysis on these data – including signature heatmaps, individual boxplots, and AUC boxplots. Interpret your findings.