Break Out Session 8

Overview

In this tutorial, we provide:

1. Six samples of paired end Bulk-RNA sequencing data.
    - **Control:**
        - Sample1_control_R1.fq
        - Sample1_control_R2.fq
    - **Treated:**
        - Sample3_treated_R1.fq
        - Sample3_treated_R2.fq

2. Two Jupyter notebooks containing the tools required to complete the exercise: One in **python** and one in **R**.

3. Two unique bioinformatics pipelines for generating gene transcripts.

Google Colab

1) create a folder in your main google drive called:
   - differential-expression-data

2) visit this link to a google colab notebook
https://colab.research.google.com/drive/19lQD1G3s35iD4YoTMgyc_dXKYZ_BZ3Ge?usp=sharing

3) Save a copy to your drive: click file then save copy in drive

Google Colab

4) **Run all the cells until you reach the cell titled:**

- ***Software Packages installation and environment setup complete.***
  - **When prompted for installing packages, type y and press enter.**
  - **Make sure to run the package installation cells in order. Wait for each one to complete and once they are done, do not rerun them**

*** Installation will take 5-7 minutes.**

Google Colab

5) Run through the cells and follow along through the collab to generate the final outputs. For the accompanying assignment, each student will be given a unique dataset to run the analysis.

** For the sake of this tutorial, this notebook simulates the reads for you.

Google Colab

6) Once you have successfully run through all of the cells contained within the first google colab file which has been configured for a python environment:

- Confirm that you have moved the gene transcript files generated by Salmon into your forked git repository.
- Then git add, commit and push those changes to your feature branch from within the google colab.

Google Colab

7) Now that you have generated the transcript quant files from salmon and pushed to your remote git repository, you are ready to pull those files into the R configured google colab for running the tool, Deseq2.
****Note, since R colab doesn't allow mounting drive or github access, the notebook comes with a gene_count matrix ready installed.**

To access the R notebook, go to the link and save a copy to your drive:
https://colab.research.google.com/drive/1KDyOC-nEjlmo-qrCtRW3qK7N05FpB2RK?usp=sharing

Then run through the cells to generate the differential analysis and figures.

.

Summary

You have now learned two ways to perform a differential expression analysis on rna sequence data.

Why choose one over the other? Some might argue that using a reference genome to align all of the reads is a more reliable approach to ensuring that all the gene transcripts are captured accurately. Ultimately, different tools will yield different results depending on their default thresholds.

The downside to using the whole genome alignment approach is that the alignment phase can take 10 or more hours to compute on the cluster. The Kmer based approach however can deliver results all the way to differential expression in under an hour. That said, often times you will need the alignment for other types of analyses on the data, so you might as well do it anyways. You'll learn about this in Part 2.

Cluster

While the google colab is remarkably convenient for running bioinformatics analyses in the cloud, it is not a scalable approach when dealing with large amounts of data. The proper way to do bioinformatics analysis is through a cluster like USC's HPC. However, since you have your own account, you will have to configure the environment and navigate the command line for this tutorial.

If you do not yet have an account or the unix and cluster skills to do so, take some time to check out the other tutorial:

git clone https://github.com/Mangul-Lab-USC/UNIX-AND-USC-HPC.git

Cluster

If you have a USC HPC account and are ready to put your unix skills to the test, then follow these instructions to complete the tutorial.

1) SSH into to your usc hpc account and navigate to your project directory.

2) Download and install your own copy of anaconda3.

- **wget https://repo.anaconda.com/archive/Anaconda3-2019.10-Linux-x86_64.sh**

3) The remaining packages for this tutorial are available through conda package manager

- **Bioconda**
- **Bioconductor**
- **Samtools**
- **Hisat2**
- **Htseq**
- **Salmon**
- **Deseq2**

Make sure you point to the conda version you just installed when installing these packages.

Cluster

4) Download the data.

git clone https://github.com/<your_account_name>/RNA-SEQ-Tutorial-PART1.git

5) Use the google collaboratory as a guide to complete the tutorial on the cluster. The commands are very similar. Simply exclude the exclamation mark (!).

6) As a challenge, write the commands inside of .sh files. Then submit those .sh files as jobs to the cluster. In real life scenarios it will take several hours for the cluster to finish computing so submitting jobs is the only feasible way to optimize your time. Again, if you are not sure how to do this, refer back to the UNIX and USC HPC tutorial mentioned in the previous slide.