

PHS 675 Special Topics: Analysis of RNA-Sequencing Data

Homework 4: Alignment and Read Counting

In this homework, you will use the same techniques from homework 3 to run hisat2 and htseq-count on all the samples.

You should submit a list of Linux commands that achieve the steps outlined below. In a narrative that follows, describe briefly what each command does. When you submit the homework, I will check your aligned, name_sorted, and counts folders to make sure everything worked.

You can refer to steps in homework 3 which I'll reference where needed, and of course refer to the course notes for each of the appropriate tools you are using. All this work should be done on biocluster, in your usual work folder, using the phs675 conda environment.

1. Create a variable that holds a list of all the sample names. (Refer to step 8 of homework 3).
2. Iterate through the list of samples and run hisat2 on each sample. Pipe the results through samtools view to create a bam file for each sample. Submit these calls to the cluster using srun. Refer to step 10 in homework 3.
3. Wait for all the alignments to complete. If you want, you can log out and log back in later when step 2 has finished (but note you will need to repeat step 1 once you have logged back in and navigated back to your folder). Once the alignments are done, iterate through all the samples and call samtools sort on each of the alignment files, sorting them by name and storing the results in a folder called name_sorted. Again, submit these calls to the cluster using srun.
4. When the sorting is complete, iterate through all the samples and call htseq-count on each of the sorted bam files, and put the output in a folder called counts. Submit these calls to the cluster using srun.