HW4 For BICB 8510

Yangyang Li
Department of Computer Science
University of Minnesota
li002252@edu.com

1 Question 1

What directory are the test data in? How many samples are we running the pipeline on? How many biological replicates are in each group?

Answer:
The directory */home/msistaff/public/RNAseq_Tutorial/Reads* contains the test data. Also, there are 2 groups including 8 samples running in the pipeline and 4 biological replicates exist in each group.

2 Question 2

What does the $CHURP represent? (hint echo $CHURP). How is this a helpful variable?

Answer:
$CHURP means that a simple way to run the program, and it actually is an alias of

```
python3 -B /panfs/roc/msisoft/churp/0.2.2-slurm/churp.py
```

3 Question 3

What role does HISAT2 play in this pipeline? If you were analyzing RNA-seq data from human samples where would you find the genome index needed for HISAT2?

Answer:
HISAT2 is mainly used to align RNA-seq data to the reference genome. For the current data analysis, we can find the genome index needed for **HISAT2** at */home/msistaff/public/RNAseq_Tutorial/Reference/*. However, you can also find the genome index of multiple organisms at the directory */panfs/roc/risdb_new/ensembl/main/*.

4 Question 4

How many fragments are in each sample? Is this enough to get a genome-wide measure of RNA expression?

Answer:
There are 8 samples and information of fragments are as follows:

Sample	R1 Fragments	R2 fragments
BoneMarrow-1	1006504	1006504
BoneMarrow-2	641889	641889
BoneMarrow-3	840077	840077
BoneMarrow-4	1044925	1044925
Spleen-1	1007898	1007898
Spleen-2	1092217	1092217
Spleen-3	982184	982184
Spleen-4	1020670	1020670

As the size of the human genome is about 3.0 Gb and the proportion of high-quality alignments that are in exons more than 0.5, I think the numbers of fragments of samples is enough to get the genome-wide measure of RNA expression.

5 Question 5

What percent of Spleen1 reads were Unmapped? (approx is ok)

Answer:
The percent of Spleen1 reads that are unmapped is about 12.85% in terms of the HTML report.

6 Question 6

Do the samples cluster as you would expect in the heatmap and MDS plot? Why would you expect that pattern?

Answer:
samples cluster in the heatmap match what I expect because samples that belong to one group are in one cluster. However, Spleen4 in the MDS plot far from other samples of its group, it may be due to some problems about sequencing or sampling.

7 Question 7

If you wanted to download the count table for all samples where is it?

Answer:
we are able to download the count table for all samples in the file *subread_counts.txt* that exists in the directory *Out/Counts*

8 Question 8

What added information do you need to calculate TPM that you don't need to calculate CPM?

Answer:
If we would like to know the TPM, we need the length of the related genes to calculate TPM compared to calculate CPM.