Homework 1

Created @Oct 8, 2020 11:30 AM

Data retrieval

Please download some RNA-seq data from the TCGA head and neck cancer project (TCGA-HNSC), explore the data, and create a histogram of the counts (from at least one) of the samples. Please write up your assignment, showing how you downloaded the data and the code to plot the histogram.

In order to retrieve the data, I went to the TCGA data repository and filtered for TCGA-HNSC, RNA-Seq, Transcriptome profiling, and HTSeq-Counts. There were in total 546 files for which I downloaded the manifest. With the following `wget` command, I download the GDC data transfer tool and subsequently, retrieved the relevant data. The following are the commands used:

```
wget <https://gdc.cancer.gov/files/public/file/gdc-client_v1.6.0_Ubuntu_x64-py3.7_0.zip>
unzip gdc-client_v1.6.0_Ubuntu_x64-py3.7_0.zip
./gdc-client download -m ./gdc_manifest.2020-10-06.txt
```

The data is organized such that each sample is a folder and the corresponding HTSeq-count data is within each folder. I chose sample `db84aaa8-8c60-44c5-b67d-65d602f9bbe9` and explored the htseq-count data `e5979bfb-70b7-4ac4-a4a0-0c045576e36d.htseq.counts`.

The following is the code to put the data into a dataframe and to plot a histogram (25 bins) of the read counts:

```
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt

%matplotlib inline

path = "./db84aaa8-8c60-44c5-b67d-65d602f9bbe9/e5979bfb-70b7-4ac4-a4a0-0c045576e36d.htseq.counts"

raw_df = pd.read_csv(path, sep="\\t", header=None, names=['gene','counts'])
cleaned_df = raw_df[raw_df['gene'].str.contains('ENSG')]

fig, ax = plt.subplots()
cleaned_df['counts'].hist(ax=ax, bins=25)
ax.set_yscale('log')
ax.set_xlabel('Read Counts')
ax.set_ylabel('Number of Genes')
```

Below are histogram plots of the data with 10 bins (top) or 25 bins (bottom):