CSE 5370: Bioinformatics
CSE 4392: Special Topics
Homework 3

Due at 4:59pm CST on Wednesday, November 9th, 2022

In this homework your will be writing code to conduct a mini Genome Wide Association Study (GWAS). The designed time to completion is 4 hours. This assignment is due at 4:59pm CST on Wednesday, November 9th, 2022.

Logistics, Expectations, & Extra Credit

Assignment Submission & Specifications

All of these files should be included in a single zip folder named "StudentLastName_UTAIDNumber_HW1_CSE5370.zip" and submitted to Canvas. It is your responsibility to ensure that files are not corrupted (It is recommended to make sure that you test your .zip files to ensure that they can be decompressed) and that your code compiles/runs. Any corrupted files or code that does not compile or run will not be given credit. Non-typeset submissions will not receive credit.

Academic Honesty & Office Hours

Many of the answers on CHEGG and similar sites that appear similar to questions on this assignment have incorrect answers. Students are encouraged to refer back to lecture recordings/slides and come to office hours before the assignment is due if they are struggling.

Group Work

Group work is explicitly allowed, however you must include a collaboration statement in your write up saying who you collaborated with and for which problems. Additionally, some coding problems require an individual submission based on a individualized data set generated randomly from your UTA ID. Every person will have their assignment graded individually.

StackOverflow.com & Similar Sites

Use of stackoverflow.com and other sites is explicitly allowed (industry researchers and academic labs use these sites frequently). However, for this course you must include a comment in your code with the link to the page you referenced whenever these sites influence your own code writing. For example, when writing this homework assignment I forgot how to insert code into LaTeXdocuments and recalled how to after visiting stackoverflow.com. If I were submitting this as an assignment, I would want to include a comment like the below example in my code submission:

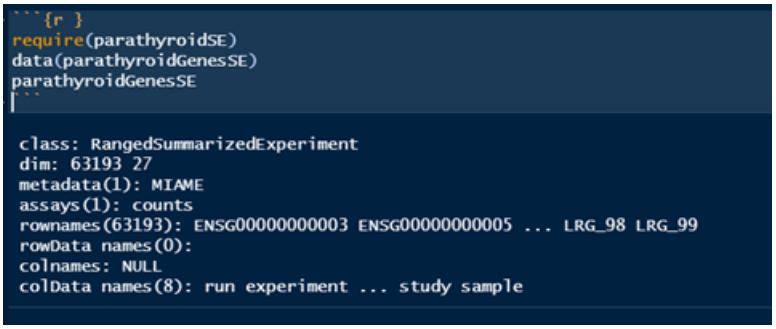
```
%When writing this homework assignment, I did not recall how to
%insert code in a nice looking way into LaTeX documents,
%so I referred to this page on stackoverflow for help:
%https://stackoverflow.com/questions/3175105
\usepackage{minted}
\begin{minted}[mathescape, linenos]{python}
Code To Insert in \LaTeX...
```

It is academic dishonesty to copy code from sites like stackoverflow without attribution like this, but is fine as long as you include attribution.

1 RNA Seq

1. In Rstudio, make a new R markdown document. Write all comments and answers to questions outside of chunks and write all code inside of chunks.

2. Install and load the `parathyroidSE` package.



This package is an example RNA seq dataset in humans with the following samples:

```
> colData(parathyroidGenesSE)
DataFrame with 27 rows and 8 columns
            run experiment  patient treatment     time submission     study    sample
    <character>   <factor> <factor>  <factor> <factor>   <factor>  <factor>  <factor>
1     SRR479052  SRX140503        1   Control      24h  SRA051611 SRP012167 SRS308865
2     SRR479053  SRX140504        1   Control      48h  SRA051611 SRP012167 SRS308866
3     SRR479054  SRX140505        1       DPN      24h  SRA051611 SRP012167 SRS308867
4     SRR479055  SRX140506        1       DPN      48h  SRA051611 SRP012167 SRS308868
5     SRR479056  SRX140507        1       OHT      24h  SRA051611 SRP012167 SRS308869
...         ...        ...      ...       ...      ...        ...       ...       ...
23    SRR479074  SRX140523        4       DPN      48h  SRA051611 SRP012167 SRS308885
24    SRR479075  SRX140523        4       DPN      48h  SRA051611 SRP012167 SRS308885
25    SRR479076  SRX140524        4       OHT      24h  SRA051611 SRP012167 SRS308886
26    SRR479077  SRX140525        4       OHT      48h  SRA051611 SRP012167 SRS308887
27    SRR479078  SRX140525        4       OHT      48h  SRA051611 SRP012167 SRS308887
```

Take the last four digits of your student ID. With each digit corresponding to a sample, subset the dataset so that it only contains the samples that are in the last four digits of your student ID. For example, if your digits are "−1129", then your dataset should contain the following samples:

```
DataFrame with 4 rows and 8 columns
            run experiment  patient treatment     time submission     study    sample
    <character>   <factor> <factor>  <factor> <factor>   <factor>  <factor>  <factor>
1     SRR479052  SRX140503        1   Control      24h  SRA051611 SRP012167 SRS308865
2     SRR479052  SRX140503        1   Control      24h  SRA051611 SRP012167 SRS308865
3     SRR479053  SRX140504        1   Control      48h  SRA051611 SRP012167 SRS308866
4     SRR479060  SRX140511        2       DPN      24h  SRA051611 SRP012167 SRS308873
```

Make a new column in the Sample table (`colData()`) called `Comparison`, and add to it the following values `c("a","a","b","b")`.

3. Get the dataset into a `DESeqDataSet` object and filter the dataset for lowly expressed genes.

4. Perform a `rlog` transformation on the dataset and generate a heatmap of the top 100 expressed genes and Principal component plot, justify why or why not each sample likely represents the new `Treatment` column that you have added and if you should throw out any samples.

5. Generate a differential expression analysis looking for differentially expressed genes in the `"b"` treatment as compared to the `"a"` treatment. How many genes are differentially expressed with a $LFC > 2$ and an FDR adjusted $p - value < 0.05$? Show the top 10 most differentially expressed genes. Plot the results in a volcano plot using the `enhancedVolcano` package.

6. Using `shinyGO`: www.bioinformatics.sdstate.edu/go, determine if there is any GO pathways that are enriched among differentially expressed genes. If there are, import any table that `shinyGO` ouputs into R studio and print it.

7. If there are differentially expressed genes and enriched pathways, explain how the original treatment and factor conditions of the samples you compared could have effected this or if there aren't, justify why there are no differentially expressed genes.

8. Finally, compile the markdown document into a PDF document and upload it to Canvas. Submit both this file and a file containing the code.

2 Difficulty Adjustment

Your answers to this section will be used to adjust the difficulty of future assignments in the class.

- How long did this assignment take you to complete?
- If the assignment took you longer than the 10 hours, which parts were overly difficult?