Bi/BE/CS183 2020
Profs. Matt Thomson and Lior Pachter
Problem Set 3
Due on Moodle by Thursday January 30th at 12:00pm

Homework policy: You may collaborate with anyone on the homework and consult/use any resources. Please write down on the homework who you worked with and explain what, if any, help you received. Also list all resources that were used for the homework and explain how they helped.

Problem 1 (10 points)

Take a look at the RNA-seqlopedia: https://rnaseq.uoregon.edu

a) What is a DNA library?

b) What is the difference between single-end and paired-end reads? When would one or the other be desirable?

c) Explain each step in an experimental RNA-Seq protocol (try to be succinct).

Note: We also uploaded to moodle the 2009 review article *RNA-Seq: a revolutionary tool for transcriptomics*, which may be helpful for understanding RNA-Seq.

Problem 2 (10 points)

Describe (in one paragraph each) the following file formats, where they come from, and what they are used for. Provide one example entry for each file (e.g. one read from some you used before, or you can write an arbitrary valid entry.

a) `FASTA`

b) `FASTQ`

c) `GFF` (or `GTF`)

d) `SAM` (and explain what is `BAM`, no need for example)

Problem 3 (25 points)

Read Lior's blog post on PCA:
https://liorpachter.wordpress.com/2014/05/26/what-is-principal-component-analysis/

a) Explain what is PCA

b) Explain what is the multi-dimensional scaling algorithm

c) Implement the multi-dimensional scaling algorithm and confirm that it works correctly by running it on a Euclidean distance matrix and comparing the output to PCA.

Problem 4 (25 points)

Read Lior's blog post describing BioJupies:
https://liorpachter.wordpress.com/2018/12/18/how-to-write-a-paper-in-four-minutes/

Try running BioJupies with any dataset (such as the one used in the blog post). Select all the available analysis options (check all the boxes). Explain (in 1-3 sentences each) what each one of the analysis performed by BioJupies is. These are enumerated below.

1. **PCA** - Linear dimensionality reduction technique to visualize similarity between samples
2. **Clustergrammer** - Interactive hierarchical clustering heatmap visualization
3. **Library Size Analysis** - Analysis of readcount distribution for the samples within the dataset
4. **Differential Expression Table** - Differential expression analysis between two groups of samples
5. **Volcano Plot** - Plot the logFC and logP values resulting from a differential expression analysis
6. **MA Plot** - Plot the logFC and average expression values resulting from a differential expression analysis
7. **Enrichr Links** - Links to enrichment analysis results of the differentially expressed genes via Enrichr
8. **Gene Ontology Enrichment Analysis** - Identifies Gene Ontology terms which are enriched in the differentially expressed genes
9. **Pathway Enrichment Analysis** - Identifies biological pathways which are enriched in the differentially expressed genes
10. **Transcription Factor Enrichment Analysis** - Identifies transcription factors whose targets are enriched in the differentially expressed genes
11. **Kinase Enrichment Analysis** - Identifies protein kinases whose substrates are enriched in the differentially expressed genes
12. **miRNA Enrichment Analysis** - Identifies miRNAs whose targets are enriched in the differentially expressed genes
13. **L1000CDS2 Query** - Identifies small molecules which mimic or reverse a given differential gene expression signature
14. **L1000FWD Query** - Projects signatures on a 2-dimensional visualization of the L1000 signature database

Problem 5 (30 points)

Read the paper *Sirt1 protects from K-Ras-driven lung carcinogenesis*, DOI 10.15252/embr.201643879 (uploaded to moodle last week).

a) Describe in one or two paragraphs the experiments done in the paper and their findings.

b) Run BioJupies GEO dataset GSE115179 (pulse experiment). Describe similarities and differences in the analyses of the paper and BioJupies.

c) Run BioJupies GEO dataset GSE115186 (pulse and chase experiment). Describe similarities and differences in the analyses of the paper and BioJupies.

d) Rerun the BioJupies analysis for experiment GSE115179 (pulse experiment) with the cases and controls are scrambled. What did you expect to find? What did you find?

e) Rerun the BioJupies analysis for experiment GSE115186 (pulse and chase experiment) with the cases and controls are scrambled. What did you expect to find? What did you find?