PUBH 7445 Final Project Instructions
October 30, 2019

Important Due Dates:
Final Project Proposal Due Date: November 27, 2019
Final Project Write-Up Due Date: December 18, 2019, 5PM

The final project provides an opportunity to conduct an in depth examination of results that have been published or to work on a dataset to which you have access. You need to submit a proposal which describes what you plan to do and where the data will come from. The proposal is not graded: the purpose is to ensure your project meets expectations. The following websites have detailed information on obtaining data from the SRA database which provides access to many next generation sequencing data sets.

`https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/`
`https://www.ncbi.nlm.nih.gov/books/NBK158898/`

There is also an R package called SRAdb that can be used to look for and download datasets. The R package GEOquery can assist with obtaining data sets from the gene expression omnibus (GEO), a public repository of microarray data.

Some hints to find data sets:

1. Read through the paper and see if a web address is given to download the data.
2. Go to the author's website. Usually the last author is the corresponding author which holds the experimental/computational lab.
3. In addition to SRA, one can check microarray databases: NCBI Gene Expression Omnibus (GEO), Stanford Microarray Database, EBIs ArrayExpress (there is an R package that can help with this), NCI's caArray etc.

The report should describe the goal of the research project, what data is available, how the raw data was processed and the data analysis plan. Conduct the analysis and describe your findings. If you are reproducing an analysis of a published paper, evaluate the extent to which your results are similar. The format should be single-spaced, font size 11 and no more than 10 pages (figures, tables and references included).

Potential (but not limited) choices of analyses to be performed in your project to improve and compare with the published paper include:

1. Preprocessing: probe level analysis, normalization, missing value imputation, gene filtering, read filtering and read trimming
2. Detect differentially expressed genes or transcripts
3. Dimension reduction and visualization
4. Gene or sample clustering
5. Classification analysis
6. Enrichment analysis (pathway analysis)

Suggested Papers for Final Project
Next-Generation Sequencing

1. (SEQC Project) SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol 2014 Sep;32(9):903-14. PMID: 25150838

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47792`

2. (SEQC Project) Wang C, Gong B, Bushel PR, Thierry-Mieg J et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol 2014 Sep;32(9):926-32. PMID: 25150839

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55347`

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47875`

3. (SEQC Project) Li S, abaj PP, Zumbo P, Sykacek P et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol 2014 Sep;32(9):888-95. PMID: 25150837

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47792`

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46876`

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5350`

4. Su Z, Fang H, Hong H, Shi L et al. An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. Genome Biol 2014 Dec 3;15(12):523.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49710`

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62564`

5. Sab A, Kress TR, Pelizzola M, de Pretis S et al. Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. Nature 2014 Jul 24;511(7510):488-92. PMID: 25043028.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51011`

6. Stoeck A, Lejnine S, Truong A, Pan L et al. Discovery of Biomarkers Predictive of GSI Response in Triple-Negative Breast Cancer and Adenoid Cystic Carcinoma. Cancer Discov 2014 Oct;4(10):1154-67. PMID: 25104330.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59810`

7. Long Q, Xu J, Osunkoya AO, Sannigrahi S et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. Cancer Res 2014 Jun 15;74(12):3228-37. PMID: 24713434.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54460`

8. Heise N, De Silva NS, Silva K, Carette A et al. Germinal center B cell maintenance and differentiation are controlled by distinct NF-B transcription factor subunits. J Exp Med 2014 Sep 22;211(10):2103-18. PMID: 25180063.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58973`

9. Lancini C, van den Berk PC, Vissers JH, Gargiulo G et al. Tight regulation of ubiquitin-mediated DNA damage response by USP3 preserves the functional integrity of hematopoietic stem cells. J Exp Med 2014 Aug 25;211(9):1759-77. PMID: 25113974.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58495`

Microarrays

1. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J et al. Gene expression in peripheral blood mononuclear cells from children with diabetes. J Clin Endocrinol Metab 2007 Sep;92(9):3705-11. PMID: 17595242.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9006`

2. Holleman A, Cheok MH, den Boer ML, Yang W et al. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. N Engl J Med 2004 Aug 5;351(6):533-42. PMID: 15295046.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE635`

3. Spira A, Beane J, Shah V, Liu G et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A 2004 Jul 6;101(27):10143-8. PMID: 15210990.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE994`

4. Gutirrez NC, Lpez-Prez R, Hernndez JM, Isidro I et al. Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. Leukemia 2005 Mar;19(3):402-9. PMID: 15674361.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1729`

5. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005 Feb 19-25;365(9460):671-9. PMID: 15721472.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034`

6. Ockenhouse CF, Hu WC, Kester KE, Cummings JF et al. Common and divergent immune response signaling pathways discovered in peripheral blood mononuclear cell gene expression patterns in presymptomatic and clinically apparent malaria. Infect Immun 2006 Oct;74(10):5561-73. PMID: 16988231.

   `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5418`