STAT718/BIOL703 Final project instructions

Final report (30% of Grade):
The format should be single-spaced, font size 12 and no more than 10 pages (figures, tables and references included). You can choose a topic and dataset that you are interested or pick one of the suggested papers on page 2-3 and follow the instructions below. Submit a proposal of your final project before class on November 11, 2024.

Important Due Dates:
Final Project Proposal Due Date: November 11, 2024 before class
Final Project Write-Up Due Date: December 11, 2024 before 5PM

Here are some loose guidelines of how you can prepare the report.

1. Data sources:
   (a) Where did you obtain the data? Are you able to obtain the raw data (e.g. CEL files) or only processed data?
2. Review of the paper:
   (a) Experimental design: Type of array? Or Type of Technology used? What samples are applied? Specific condition, mutation or treatments? Biological replicates or experimental replicates?
   (b) Data preprocessing: How are the data preprocessed? Any normalization and filtering performed?
   (c) Data analysis: What analyses are performed in the paper?
   (d) Conclusion: Describe the conclusions in the paper. Any further validation experiments performed? What is the implication and biological importance of the study?
3. Evaluating and re-analyzing the data:
   (a) Try your best to repeat the analytical procedures in the paper. If the description in the paper is not clear or it involves complicated methods not taught in class, try to use similar analyses.
   (b) Comment on the pros and cons of their analyses and perform better alternatives based on what you have learned in this class. Do you obtain improved results? (Most of the papers were published early before 2002 and many new methods may actually improve their results.)
   (c) Any other further analysis? Report what you did and interpret your result and its implications.
   (d) It is suggested to collect another array study of the same disease and perform some sort of meta-analysis with your data. If meta-analysis does not improve,

discuss the potential reasons of heterogeneity.

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

(1) Preprocessing: probe level analysis, normalization, missing value imputation, gene filtering.
(2) Detect differentially expressed genes
(3) Dimension reduction and visualization
(4) Gene or sample clustering
(5) Classification analysis
(6) Enrichment analysis (pathway analysis)
(7) Other advanced co-regulation analysis
(8) Genomic meta-analysis

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 who holds the experimental/computational lab.
3. Google.
4. Check microarray databases: NCBI Gene Expression Omnibus (GEO), Stanford Microarray Database, EBI's ArrayExpress, NCI's caArray, Sequence Read Archive (SRA) etc.

Suggested Papers for Final Project

Next-Generation Sequencing

Single-Cell RNA Sequencing

1. Hou, W., Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis.Nat Methods 21, 1462–1465 (2024). https://doi.org/10.1038/s41592-024-02235-4
2. He, S., Bhatt, R., Brown, C. et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nat Biotechnol 40, 1794–1806 (2022). https://doi.org/10.1038/s41587-022-01483-z
3. Zhicheng Ji, Hongkai Ji, TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, *Nucleic Acids Research*, Volume 44, Issue 13, 27 July 2016, Page e117, https://doi.org/10.1093/nar/gkw430

4. Kakaradov B, Arsenio J, Widjaja CE, He Z et al. Early transcriptional and epigenetic regulation of CD8 T cell differentiation revealed by single-cell RNA sequencing. Nat Immunol 2017 Apr;18(4):422-432. PMID: 28218746
5. Deng J, Wang ES, Jenkins RW, Li S et al. CDK4/6 Inhibition Augments Antitumor Immunity by Enhancing T-cell Activation. Cancer Discov 2018 Feb;8(2):216-233. PMID: 29101163
6. Zemmour D, Zilionis R, Kiner E, Klein AM et al. Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. Nat Immunol 2018 Mar;19(3):291-301. PMID: 29434354
7. Ito Y, Ashenberg O, Pyrdol J, Luoma AM et al. Rapid CLIP dissociation from MHC II promotes an unusual antigen presentation pathway in autoimmunity. J Exp Med 2018 Oct 1;215(10):2617-2635. PMID: 30185635
8. Gutierrez-Arcelus M, Teslovich N, Mola AR, Polidoro RB et al. Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. Nat Commun 2019 Feb 8;10(1):687. PMID: 30737409
9. Huang H, Sikora MJ, Islam S, Chowdhury RR et al. Select sequencing of clonally expanded CD8⁺ T cells reveals limits to clonal expansion. Proc Natl Acad Sci U S A 2019 Apr 30;116(18):8995-9001. PMID: 30992377
10. Yao C, Sun HW, Lacey NE, Ji Y et al. Single-cell RNA-seq reveals TOX as a key regulator of CD8⁺ T cell persistence in chronic infection. Nat Immunol 2019 Jul;20(7):890-901. PMID: 31209400 (Chip-Seq data is also available).
11. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C et al. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell 2018 Nov 1;175(4):984-997.e24. PMID: 30388455

RNA Sequencing

1. McCabe MS, Waters SM, Morris DG, et al. RNA-seq analysis of differential gene expression in liver from lactating dairy cows divergent in negative energy balance., BMC Genomics (2012) 20;13:193. (cow)
2. Sabò 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. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51011 [BED files]
3. 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. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59810

4. 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. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54460
5. Heise N, De Silva NS, Silva K, Carette A et al. Germinal center B cell maintenance and differentiation are controlled by distinct NF-κB transcription factor subunits. J Exp Med 2014 Sep 22;211(10):2103-18. PMID: 25180063. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58973
6. 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. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58495
7. (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
8. (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
9. (SEQC Project) Li S, Łabaj 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
10. (SEQC Project) Munro SA, Lund SP, Pine PS, Binder H et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat Commun 2014 Sep 25;5:5125. PMID: 25254650
11. 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.