STAT718/BIOL703 Final Project Instructions

Final report (30% of Grade)/Final Project Presentation (15%):
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 March 30, 2026.

Important Due Dates:
Final Project **Proposal** Due Date**: March 30, 2026 before class**
Final Project **Presentation: April 20, April 22, April 27 in class**
Final Project **Write-Up** Due Date**: May 04, 2026 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.

If you are looking for ideas or papers, the link below includes many papers that use interpretable or explainable machine learning methods in genomic data analysis. These are described in the following paper: Van Hilten, A., Katz, S., Saccenti, E., Niessen, W. J., \& Roshchupkin, G. V. (2024). Designing interpretable deep learning applications for functional genomics: a quantitative analysis. Briefings in Bioinformatics, 25(5).

https://roshchupkin.notion.site/4cbd73a2ecf542c383b1d05865205bc4?v=84f991e732c941a18bf850a579fe8c5d