PubH 8445: Final Project Instructions & Guidelines

Fall 2023

The final projects are to conduct a genomics data analysis to

1. *demonstrate your knowledge of common statistical methods for genetic/genomics data analysis*
2. *explore the recent advances in methodology development for genetic/genomics data analysis*

The projects are to be done individually. You may choose your own topic for the final project, which can be related to the topics we covered/will cover in the course (GWAS, polygenic score, knock-off FDR control, Mendelian randomization, and spatial transcriptomics) or an avenue that was un-addressed in class. The sequence of stages in the project process are

**Project Topic due Friday Nov. 3 via email**

You should decide a topic for your final project by November 3. There are two directions for the projects that you can consider:

1. Comparing existing methods for analyzing existing/simulated genetic/genomics data;

2. Replicating the work of recent methodology papers in genetic/genomics data analysis.

For Option 1, the links for potential data sources are provided at the end of the document. For Option 2, the instructors will provide a list of candidate papers on canvas. You are also welcome to investigate and/or use datasets or papers beyond the provided lists, which can be related to your research area (if you have one). Do not submit anything you have completed prior to attending the course.

**Project Proposal due Friday Nov. 17 noon via email**

To judge the suitability of your choice of a project, you should produce a 1-2 page Project Proposal first that describes your idea for the project and the work you intend to perform. This proposal should outline:

1. A high-level statement of the problem you intend to address and the goals of your project

2. The data source(s) you intended to use

3. Description of the analysis methods to be used/examined

4. The results you plan to present, ideally including plots for data visualizations and result presentation, and metrics for method/data analysis assessment.

5. Expected results and potential challenges

The proposal will then be assigned to others for peer review.

**Peer Review of Project Proposal due Tuesday Nov. 21 in class**

Each student will be assigned a Project Proposal written by another student. The reviewer should write a report to provide positive feedback as well as constructive criticism for the reviewed proposal. You may also ask critical questions for the author to address. There will be an in-class discussion session during which the reviewer and reviewee will discuss over the review reports.

**Project Presentation Tuesday Dec. 5 in class**

You need to prepare a PowerPoint presentation of around 15-20 min. The presentation should include

1. Problem statement: What goal of the project?
2. Data: What type of analysis will you conduct and what dataset you will use?
3. Methods you used and the rational for their use: May include data preparation/preprocessing, method evaluating metrics, etc.
4. Results (may be preliminary): Any results you have to report so far. May also report unexpected challenges.
5. Conclusions and comments.

**Final Report due Friday Dec. 15 via email**

The project report is the formal description of your project. The report should be 5-10 pages in length (preferably using Tex). The structure is similar to the presentation but with full elaboration on what you did.

1. Problem statement and background

Give a clear and complete statement of the problem.

Include background material as appropriate: who cares about this problem, what impact it has, what implications better solutions might have.

Included a brief summary of any related work you know about.

Overview the statistical methods that you plan to use and justify your methods in terms of the problem statement.

Include measures that you plan to use for model assessment/selection.

2. Data

Describe in detail the datasets that you used, including the sources of data, dimension and characteristics of the data.

Provide the steps for data preprocessing and/or assembly as appropriate.

3. Methods

Describe the statistical methods that you used in detail. Justify your choice of the methods in terms of their features, including the pros/cons of each method.

Describe the computational tools/software/packages that you use for the methods.

4. Results

Give a detailed summary of the results of your work. Present the data analysis results and the measures that you used for model assessment/selection.

Present your findings from the analysis in terms of your goal of the project.

Please use tables and figures to display your results whenever possible.

5. Conclusion/Discussion

Elaborate discoveries, future directions, or any other comments.

**Grading policy**

Project Proposal 15pt

Peer Review Report 10pt

Project Presentation 30pt

Questions during Presentation 5pt

Final Report 40pt

Total 100pt

**Potential Data Resources;**

1. **Genetic data (for GWAS, polygenic score, genetic feature selection, knock-off FDR control)**

NCBI Database of Genotypes and Phenotypes:

<https://www.ncbi.nlm.nih.gov/gap>

UK Biobank:

<https://www.ukbiobank.ac.uk/>

Pan-UK Biobank

<https://pan.ukbb.broadinstitute.org/>

Health and Retirement Study (HRS)

<http://hrsonline.isr.umich.edu>

Wisconsin Longitudinal Study (WLS)

<https://www.ssc.wisc.edu/wlsresearch/>

LifeLines Biobank in the Netherlands

<https://www.lifelines.nl>

GTEx:

<https://gtexportal.org/home/>

1. **GWAS summary statistics data (for meta-analysis, polygenic score, multi-trait analysis, Mendelian randomization)**

The NHGRI-EBI GWAS Catalog:

<https://www.ebi.ac.uk/gwas/summary-statistics>

Atlas of GWAS Summary Statistics:

<https://atlas.ctglab.nl/>

Dr. Neale’s GWAS summary statistics from UK Biobank

<http://www.nealelab.is/uk-biobank/>

MR Base for Mendelian Randomization tutorial:

<https://www.mrbase.org/>

Other resources: Genomics Data Lake

<https://learn.microsoft.com/en-us/azure/open-datasets/dataset-genomics-data-lake>

1. **Single-cell/Spatial transcriptomics (for spatial genetic/genomics data analysis)**

GTEx:

<https://gtexportal.org/home/>

10x Genomics:

<https://www.10xgenomics.com/resources/datasets>

Vizgen:

<https://vizgen.com/data-release-program/>