Faculty of Life
Sciences
Systems Genetics 2023

Final Project
Guidelines

Systems genetics 2023
Due: August 31, 2023

The goal of this project is to give you the opportunity to apply the techniques that you have learned in class for combining genetics and intermediate phenotypes to discover novel findings. Submit the project in **pairs.**

The project consists of the following parts:

1. **Defining the task**: Define the task you are trying to solve clearly. The project is focused on BXD data and is meant to involve a combination of two or more intermediate phenotype collections, in addition to genotyping data. You should test your approach on a collection of physiological phenotypes of interest. For example, you may focus on the collection of behavioral traits (e.g., open field behavior in response to morphine injection), and use gene expression from both brain and liver tissues as relevant intermediate traits (since liver is responsible for the metabolism of this drug).

Some optional gene expression datasets:

Tissue / Cell type	Database	Access ID
HSC	GEO	GDS1077
Lung	Array express	E-MTAB-848
Eye	In moodle (eye_expression.zip)	
Myeloid	GEO	GSE18067
Blood stem cell	GEO	GSE18067
Erythroid	GEO	GSE18067
Hypothalamus	GEO	GSE36674
Kidney	GEO	GSE8356
Liver	GEO	GSE17522
T cells	Array express	E-MTAB-836

2. **Gene expression data preprocessing (15%):**

- Make sure you download the **normalized** data.
- Merge data file with annotation file to get your input matrix. In your input matrix, rows should be annotated with gene identifier (either gene symbol or

entrez IDs); columns are BXD strain names.

- Remove rows with no gene identifier.
- Remove rows with low maximal value (choose a threshold).
- Remove rows with low variance (choose a threshold).
- The data may contain multiple rows (probes) for the same gene. To have a single row per gene, either select one probe while removing the others (e.g., select the highest- variance probe) or calculate their average.
- In addition, similarly to EX3, filter neighboring loci.

3. **eQTL analysis (15%)**: As in assignment 3, run an association test (either ANOVA or regression) on all genes using each of the SNPs in the genotype file. For each gene, select those SNPs that are best associated with the expression of the gene (genes with weak associations for all SNPs should be excluded). The significantly associated SNPs are called eQTLs. Explain the way you chose the final collection of eQTLs and discuss their statistical significance (use multiple testing correction). Report your results (for example, how many different eQTLs? What is the distribution of number of genes associated with a given eQTL?)
4. **QTL analysis (15%):** As in assignment 2, run genome wide association test on each of the selected phenotypes. Each significantly associated SNP is called QTL. Discuss the statistical significance of the identified QTLs. Pay attention to multiple testing correction.
5. **Combine results (15%):** Compare the QTLs of your phenotypes with the collection of eQTLs. Does gene expression data provide any added value for the identification of QTLs? From your experience, what conclusion can be drawn about limiting GWAS to those DNA variants that are associated with at least one expression trait?
6. **Causality analysis (40%):** Apply the causality test on the results from previous parts. Run causality test on each pair of gene and phenotype where both the QTL and the eQTL are located in a nearby genomic position (or, of course, have the same position). Report the predicted relations among the QTL, associated gene, and phenotype. Apply permutation test to get statistical significance for one or a few specific causality hypotheses. Run the test on 10 triplets and explain the choice you made and the design of permutation test in detail.

Submission Guidelines:

1) Submit a zip file named final_project_<your_id>_<your_partner_id>.
2) Provide your code, report and everything you find necessary.
3) Add the results of your code to the report, don't make us run it.
4) Elaborate about the decisions you're making throughout the way.
5) Data: mention the data you're using. If you chose a different dataset than those mentioned above, provide a link to it.
6) We advise you to submit a clean and organized code. In case of wrong answers, it will assist us in finding the cause and reduce points deduction.