Mini Project 1

Spring 2020

Prof. Alexis Battle

DUE DATE: 03/10/2020

1. Writing introductory text.

- For each of two papers we have read in class (any paper we read up through the due date), you will write alternative introductory text. This should be 6-10 sentences that would begin the introduction (**NOT the abstract**), introducing the topic and importance of the paper.
    - Select a paper from the readings that, as written, appears to target a broad rather than highly specialized audience. Write new introductory sentences as if you were targeting more specialized experts in the topic of the paper.
    - Select a second paper that appears to target a specialized audience. Write new introductory sentences as if you were targeting a broad audience, such as Science/Nature/grant reviewers.

2. Paper critique

- For one of the papers we have read up through the due date, write an approximately one page review of the paper, as if you were reviewing the paper for a journal. Specifically note any problems with clarity of the text; potential problems, confounders, or simply questions about the analysis; and additional experiments you would like to see. Assess the overall impact of the work briefly.
    - Begin with a brief summary of the paper and its overall importance
    - Go on to note specific questions, concerns, and requests

3. Data analysis:

Here, you will use data from the GTEx project.

Please refer to the GTEx supplement and documentation on the GTEx portal with initial questions, though you can also work together, post questions on slack, and ask me for clarification as needed.

- Run eQTL analysis for GTEx data
- Explore multiple hypothesis correction approaches for eQTL analysis
- Compare strategies for correcting latent confounders in GTEx tissues

You will be graded on the correctness of your analysis, reproducibility, and thought you put into evaluating your approach.

Data can be found on MARCC under:
/work-zfs/abattle4/lab_data/GTEx_v8_eqtl_practice/matrix_eqtl

NOTE: please perform all analysis on MARCC – genotype data cannot be stored on your local laptop / device.

Your tasks:

1. You will need genotype (chr 10), covariate, and gene expression files for Blood tissue. Make sure you look at the files and understand their contents.

2. Run PCA on provided gene expression data to estimate latent confounders
    a. Save the output (both components and loadings)
    b. Make scree plot and include in your submission

3. Call cis-eQTLs from Blood, chromosome 10, using only *genotype* PCs and sex as covariates (found in the "covariates" file). **You can use software packages Matrix eQTL or FastQTL (files are formatted for Matrix eQTL)**
    a. How many eQTLs do you get if you apply Benjamini-Hochberg with FDR threshold 0.05 across all tested all SNP-gene pairs (NOT gene level FDR)?
    b. How many eQTLs do you get if you perform **gene-level FDR** control at 0.05. To obtain gene level p-values, you can perform permutations or Bonferroni correction within each gene as discussed in lecture.

4. Include 0, 5, 10, 20, and 30 expression PCs (along with sex and genotype PCs) as covariates during eQTL calling, rerunning eQTL analysis for chromosome 10 for each setting, using *gene level* FDR 0.05 as a significance cutoff (with the same approach you chose in 3)
    a. Plot the number of genes with an eQTL you obtain vs number of PCs used in your analysis and include in your submission

5. Considering the scree plot and the results of # 4, how many factors do you think you should regress out from each tissue for optimal eQTL calling? Would you run any other analysis before deciding? Justify your answer.

6. Submit all plots, answers to above questions, and scripts you used to run each step. Please use R or python.