The advent of whole-genome sequencing technologies has allowed for the sequencing of countless organisms. Importantly, the cost associated with whole-genome sequencing has dropped greatly over the last decade1. This reduction in cost was met with an increase in genome-wide association studies (GWAS) in humans and a variety of other model organisms2,3. Similarly, there has been a rise in linkage analysis4. Both GWA and linkage mapping aim to link phenotypic differences in a test population to genomic differences, which are called quantitative trait loci (QTL). The difference in the techniques is centered around the difference in test subject, linkage analysis requires individuals to be direct descendants of a common parent, while GWA analysis does not. These are powerful methods for identifying regions of the genome that are associated with a particular phenotypic trait, a certain disease for example. Recently, researchers realized that traditional mapping methodologies can be used toward studying expression levels of all genes in an organism simultaneously, a technique that is referred to as *e*QTL mapping5.

Countless dollars have been invested in the implementation of these studies, but thus far we have only attained understanding of “low-hanging” fruit (traits that are mendelian in nature). Most phenotypic traits however, including diseases, are complex in nature meaning that multiple genes, the environment, and gene-environment interactions all contribute to the phenotype. More statistical power is required to identify complex traits because when more interacting components contribute to a phenotype they each explain less of the variance associated with a trait and more statistical power is required to identify causal loci. This makes studying complex traits both difficult and costly in humans.

*Caenorhabditis elegans* is a free-living nematode that was established as a model organism in 1974 by Sydney Brenner6. Research in *C. elegans* has led to the discovery of numerous signaling pathways present in metazoans and remains a powerful metazoan model due to its simplicity of use, its low cost in maintenance, and its genetic tractability. There have been numerous genetic mapping studies performed in *C. elegans*7*,* which have the added benefit of being able to perform follow-up experiments to identify molecular mechanisms associated with causal polymorphisms.

The focus of my final project will be on the analysis presented in Francesconi *et al.*8. This study incorporates gene expression data from two separate publications9,10 to analyze the dynamics of gene expression through the developmental cycle of *C. elegans.* I will attempt to replicate the statistical analysis presented in the paper as well as seek to find holes or improvements on their analyses.

Francesconi *et al.*8 attempt to analyze how temporal gene expression dynamics are controlled by natural genetic variation present in a set of recombinant inbred advanced intercross lines (RIAILs)9. The authors performed canonical correlation analysis (CCA) on gene expression data obtained by Reinke *et al.*10*.* This data set contains microarray data collected every four hours in triplicates for 12 hours. This data set was used to identify gene expression patterns that are indicative of the animals age (canonical variates). The canonical variates were then used to identify the life-cycle stage the RIAIL samples were in.

The authors at this point have expression data for 206 RIAILs that have been assigned a rank life-cycle stage based on a reference gene expression time course. They used this data set to then perform eQTL mapping to identify genetic variants that lead to altered gene expression.

It is unclear how the authors take account for differences in gene expression levels solely based on genetic background. The authors identify 2985 cis-eQTL at .1 FDR when not taking temporal position into account, when the authors from the publication in which these data were taken from observed 2309 cis-eQTL at .05 FDR9. The difference in number is likely due to the FDR threshold set. The authors then proceed to take developmental stage into account and identify a total of 4246 cis-eQTL. However, there is no mention of incorporating differences based on genetics as a covariate into the temporal analysis. It makes sense that genetic differences should be taken into account when identifying QTL based on developmental timing because genetic differences could be leading to altered gene expression during different developmental stages.

I propose to do the following analysis:

1. Canonical correlation analysis on data generated by Reinke *et al.*10 to identify gene expression patterns that correspond to develepmental stages.
2. cis-eQTL mapping based on gene expression data from data generated by Rockman *et al.*9and see how it compares to the study of interest8. This will be done without accounting for temporal dynamics.
3. Perform #2 while taking temporal dynamics into account as is performed in the study.
4. Introduce gene expression levels based on genetic background as a covariate in for the analysis in #3.

I expect the number of QTL identified by the researchers in #3 to be significantly reduced by incorporating gene expression differences based on genetic background as a covariate in the model. I think this reanalysis will be both useful for my knowledge of statistical computing in R, challenging claims being made in a publication, and increase my knowledge of QTL mapping.

1. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).

2. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6,** 95–108 (2005).

3. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (2007).

4. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58,** 1347–63 (1996).

5. Rockman, M. V & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7,** 862–72 (2006).

6. Brenner, S. Caenorhabditis elegans. 71–94 (1974).

7. Snoek, L. B. *et al.* WormQTL--public archive and analysis web portal for natural variation data in Caenorhabditis spp. *Nucleic Acids Res.* **41,** D738–43 (2013).

8. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505,** 208–11 (2014).

9. Rockman, M. V, Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* **330,** 372–6 (2010).

10. Reinke, V., Gil, I. S., Ward, S. & Kazmer, K. Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development* **131,** 311–23 (2004).