BIOS 731 Final Project Report

Simulation of Correlated Single-Cell RNA-Seq Data and Evaluation of Co-Expression eQTL Mapping Methods

Yutong Liu

**Abstract**

Mapping co-expression quantitative trait loci (co-eQTLs) can deepen our understanding of gene regulatory networks and the genetic basis of complex diseases. While traditional expression QTL (eQTL) studies have provided important insights, they often miss the nuanced regulatory effects that manifest as altered gene co-expression. In this study, we evaluate two co-eQTL mapping methods—the Spearman correlation-based approach and the Memento moment-based approach—using simulated single-cell RNA sequencing (scRNA-seq) data that reflect realistic properties derived from the OneK1K dataset. Our simulations model co-expression as a function of genotype, capturing individual- and cell-level variability through a hierarchical Poisson-Gamma framework. We assess each method’s type I error control and statistical power across a range of gene expression levels. Our results reveal that while both methods demonstrate acceptable type I error rates at moderate expression levels, they suffer from reduced power as expression declines. The Memento method is overly conservative, especially for lowly expressed genes, whereas the Spearman method displays inflated or deflated type I errors depending on expression levels and cell sampling variability. These findings highlight key limitations of current co-eQTL methods and underscore the need for more robust statistical approaches that can account for scRNA-seq data sparsity and complex expression distributions.

**Introduction**

Understanding the genetic risk factors and mechanisms underlying complex diseases has been a longstanding goal since the early development of genetics. The advent of high-throughput sequencing technologies has enabled genome-wide association studies (GWAS) to systematically identify single nucleotide polymorphisms (SNPs) associated with disease traits. GWAS has made significant contributions, for example, the identification of SNPs near *IL28B* that facilitated the development of personalized treatments for Hepatitis C.(Balagopal et al., 2010) To date, the GWAS Catalog has reported over 799,200 associations across more than 7,000 publications. (Cerezo et al., 2024)

However, most GWAS variants are hard to interpret. First, linkage disequilibrium (LD) complicates the identification of truly causal variants. Moreover, complex diseases often involve interactions across multiple cell types, making it difficult to determine the primary cellular contexts in which risk variants exert their effects. Adding to the complexity, approximately 90% of GWAS-identified variants fall within non-coding regions of the genome, where their functional consequences are less direct. (Cano-Gamez & Trynka, 2020)

While fine-mapping techniques can address LD to some extent (Schaid et al., 2018), a deeper understanding of disease mechanisms requires investigation beyond associations. Specifically, it necessitates tracing how genetic variants influence intermediate molecular phenotypes such as DNA methylation, histone modifications, chromatin accessibility, RNA transcription and stability, protein abundance, and isoform usage. These processes are studied through regulatory quantitative trait loci (QTL) mapping. (Umans et al., 2021)

Current research on regulatory QTLs has predominantly focused on expression QTLs (eQTLs), which quantify how genetic variants affect gene expression levels. Over 4.2 million eQTL associations have been identified, comprising both cis-eQTLs, which regulate nearby genes, and trans-eQTLs, which influence distant genes. (Umans et al., 2021) While cis-eQTLs are generally easier to detect and interpret, they account for only a small fraction of expression heritability. In contrast, up to 70% of inter-individual variance in gene expression is thought to result from trans effects, which remain incompletely characterized. (Umans et al., 2021) Furthermore, eQTL effects often manifest within regulatory networks rather than as isolated gene-variant relationships; variants acting as trans-eQTLs frequently also display cis-regulatory effects that mediate their broader influence. (Umans et al., 2021) (Võsa et al., 2021)

Recognizing this network-level complexity, recent studies have sought to use single cell RNA-Seq data to construct personalized gene regulatory networks to better capture the interactions among genes influenced by genetic variants. (van der Wijst et al., 2018) Therefore, beyond traditional eQTLs, attention has turned toward variants that modify the co-expression between pairs of genes. (Li et al., 2023) Two primary methods have been developed for co-eQTL mapping: one leveraging Spearman correlation (Li et al., 2023) and another using a moment-based approach (Kim et al., 2024) However, their performance in detecting co-eQTLs remains insufficiently explored.

This study aims to evaluate the existing co-eQTL mapping methods using simulated single cell RNA-Seq data. Specifically, I will assess whether current approaches accurately detect co-eQTLs and identify their limitations in capturing co-expression signals influenced by genetic variation.

**Methods**

*Simulation Framework*

To simulate two genes whose co-expression is modulated by genotype, we borrow information, including mean expression level, number of cells in each individual, sequencing depth, from a OneK1K, which contains 981 individuals and 289,000 CD4 TCM cells. (Yazar et al., 2022)

The genotype is simulated by binomial distribution, and the co-expressed two genes are simulated based on a hierarchical model.

With in the model, zi,j stands for the latent expression level for ith individual jth gene, and the latent expression level for specific cell zic,j is conditional on the latent expression value for the individual, and the observed UMI counts xic,j is conditional on the cell specific latent expression level and effected by the sequencing depth sic of the cell. The genotype module the co-expression by value of and , and the bivariate Gamma distribution with a specified correlation is simulated using a copula-based approach, which leverages the uniform property of the inverse cumulative distribution function (CDF). (Sun et al., 2021) and are adjusted to evaluate the performance of co-eQTL methods across different expression levels. Specifically, expression levels are chosen to match those of the 100th highest expressed gene, as well as the 200th, 400th, 800th, 1600th, up to the 3200th. For each of these scenarios, 1,000 datasets are simulated.

*Methods Comparison*

There are two existing methods, the Spearman-correlation-based method (Li et al., 2023) and a Moment-based method (Memento) (Kim et al., 2024), and both are two-step method to identify co-eQTL.

The Spearman-based method first calculates the Spearman correlation coefficient ρi between two genes for each individual. It then tests for co-eQTLs by evaluating the association between ρi and genotype.

The Memento method assumes that scRNA-seq data arise from a hypergeometric sampling process, in which a subset of mRNA molecules is captured from the total mRNA pool in each cell, reflecting sequencing efficiency. Using this assumption, it estimates ρi based on the moments of the distribution. Co-eQTL associations are then tested via a linear model linking ρi and genotype, with significance assessed through bootstrap resampling.

The Spearman-based method is implemented in R 4.3.3, and the Memento method is implemented in Python 3.13 using the provided package “memento-de.”

The two methods are compared in terms of type I error and statistical power for identifying co-eQTLs. Results are visualized using bar plots with Monte Carlo error bars. Under the null setting, QQ plots are also examined to assess the uniformity of p-values and detect any inflation or deflation, which would indicate potential issues with type I error control.

**Results**

Under the null setting, Memento shows a type I error rate of 0.065 (SE: 0.0056) at the 100th expression level, exceeding the nominal 0.05 threshold. At lower expression levels, its type I error drops well below 0.05, indicating conservative behavior (Figure 1a). The Spearman-based method, in contrast, has a type I error rate of 0.03 (SE: 0.0055) at the 100th expression level—below the nominal threshold—and stays close to 0.05 for lower expression levels (Figure 1a). These patterns are also evident in the QQ plots: for Memento, observed p-values tend to exceed the expected values at lower expression levels, suggesting conservative behavior. For the Spearman method, the QQ plot at the 100th expression level shows a biphasic pattern—initially with observed p-values smaller than expected, followed by larger-than-expected values. At the 200th and 400th expression levels, the QQ plots show conservative behavior in the tails, where observed p-values exceed expected ones (Figure 2).

**a b**

A graph with blue and red bars

AI-generated content may be incorrect. A graph of a number of people

AI-generated content may be incorrect.

**Figure 1 a)** Type I error for Memento and Spearman correlation method in identifying co-expression eQTL. **b)** Power for Memento and Spearman correlation method in identifying co-expression eQTL.

Both Memento and the Spearman-based method show limited performance in terms of power for identifying co-eQTLs, with power decreasing as expression levels decline. The highest power is achieved by the Spearman method at the 100th expression level, reaching 0.942 (SE: 0.009). In comparison, Memento performs substantially worse, with its highest power reaching only 0.317 (SE: 0.013). (Figure 1b)

**A group of graphs showing different types of numbers

AI-generated content may be incorrect.**

**Figure 2** QQ plot for Memento and Spearman correlation method p values under different expression level.

**Discussion**

Co-expression eQTL (co-eQTL) mapping is an emerging area of interest in understanding gene regulatory networks and advancing disease treatment. While current methods have identified potential co-eQTLs—such as rs1131017, which influences RPS26 expression and has been associated with autoimmune diseases (Li et al., 2023) —our simulation results highlight notable limitations in their accuracy and power. The Memento method tends to be overly conservative, particularly at lower expression levels, likely due to its strong distributional assumptions. In contrast, the Spearman-based method is prone to biased estimates, especially at higher expression levels, where limited cell sampling exacerbates the issue. Additionally, both methods struggle with low expression levels due to the inherent sparsity of scRNA-seq data, which hinders reliable detection of co-expression.

To address these challenges, we propose developing a generalized estimating equations (GEE)-based approach for co-eQTL mapping. Unlike the existing two-step methods, our proposed method integrates estimation and testing into a unified framework, potentially offering greater accuracy and statistical power by making more efficient use of the data.

**GitHub:** https://github.com/ytliu36/bios731\_final\_liu.git

**References**

Balagopal, A., Thomas, D. L., & Thio, C. L. (2010). IL28B and the control of hepatitis C virus infection. *Gastroenterology*, *139*(6), 1865-1876. <https://doi.org/10.1053/j.gastro.2010.10.004>

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in genetics*, *11*, 424. <https://doi.org/10.3389/fgene.2020.00424>

Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, Karatuğ O., Hall, P., Hayhurst, J., John, S., Mosaku, A., Ramachandran, S., Foreman, A., Ibrahim, A., McLaughlin, J., Pendlington, Z., Stefancsik, R., Lambert, S. A., McMahon, A., Morales, J., Keane, T., Inouye, M., Parkinson, H., & Harris, L. W. (2024). The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research*, *53*(D1), D998-D1005. <https://doi.org/10.1093/nar/gkae1070>

Kim, M. C., Gate, R., Lee, D. S., Tolopko, A., Lu, A., Gordon, E., Shifrut, E., Garcia-Nieto, P. E., Marson, A., Ntranos, V., & Ye, C. J. (2024). Method of moments framework for differential expression analysis of single-cell RNA sequencing data. *Cell*, *187*(22), 6393-6410.e6316. <https://doi.org/10.1016/j.cell.2024.09.044>

Li, S., Schmid, K. T., de Vries, D. H., Korshevniuk, M., Losert, C., Oelen, R., van Blokland, I. V., Groot, H. E., Swertz, M. A., van der Harst, P., Westra, H.-J., van der Wijst, M. G. P., Heinig, M., Franke, L., & Bios Consortium, s.-e. C. (2023). Identification of genetic variants that impact gene co-expression relationships using large-scale single-cell data. *Genome Biology*, *24*(1), 80. <https://doi.org/10.1186/s13059-023-02897-x>

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491-504. <https://doi.org/10.1038/s41576-018-0016-z>

Sun, T., Song, D., Li, W. V., & Li, J. J. (2021). scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, *22*(1), 163. <https://doi.org/10.1186/s13059-021-02367-2>

Umans, B. D., Battle, A., & Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends Genet*, *37*(2), 109-124. <https://doi.org/10.1016/j.tig.2020.08.009>

van der Wijst, M. G. P., de Vries, D. H., Brugge, H., Westra, H.-J., & Franke, L. (2018). An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine*, *10*(1), 96. <https://doi.org/10.1186/s13073-018-0608-4>

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M.-J., Agbessi, M., Christiansen, M. W., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez Flitman, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg, J., Kettunen, J., Lee, B., Zhang, F., Qi, T., Hernandez, J. A., Arindrarto, W., Beutner, F., ’t Hoen, P. A. C., van Meurs, J., van Dongen, J., van Iterson, M., Swertz, M. A., Jan Bonder, M., Dmitrieva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Hewitt, A. W., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M. G., Penninx, B. W. J. H., Pritchard, J. K., Raitakari, O. T., Rotzschke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., ’t Hoen, P. A. C., Thiery, J., Tönjes, A., van Dongen, J., van Iterson, M., Veldink, J. H., Völker, U., Warmerdam, R., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., van Meurs, J., Prokisch, H., Ahsan, H., Pierce, B. L., Lehtimäki, T., Boomsma, D. I., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W. H., Downes, K., Stegle, O., Battle, A., Visscher, P. M., Yang, J., Scholz, M., Powell, J., Gibson, G., Esko, T., Franke, L., Consortium, B., & i, Q. T. L. C. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, *53*(9), 1300-1310. <https://doi.org/10.1038/s41588-021-00913-z>

Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R. P., Clarke, L., Maccora, K., Chen, C., Cook, A. L., Ye, C. J., Fairfax, K. A., Hewitt, A. W., & Powell, J. E. (2022). Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, *376*(6589), eabf3041. <https://doi.org/10.1126/science.abf3041>