Spring 2021

Improving Risk Factor Identification of Human Complex Traits in Omics Data

Weimiao Wu
Yale University Graduate School of Arts and Sciences, weimiao.wu@mail.harvard.edu

Abstract

Improving Risk Factor Identification of Human Complex Traits in Omics Data

Weimiao Wu

2021

With recent advances in various high throughput technologies, the rise of omics data offers a promise of personalized health care with its potential to expand both the depth and the width of the identification of risk factors that are associated with human complex traits. In genomics, the introduction of repeated measures and the increased sequencing depth provides an opportunity for deeper investigation of disease dynamics for patients. In transcriptomics, high throughput single-cell assays provide cellular level gene expression depicting cell-to-cell heterogeneity. The cell-level resolution of gene expression data brought the opportunities to promote our understanding of cell function, disease pathogenesis, and treatment response for more precise therapeutic development. Along with these advances are the challenges posed by the increasingly complicated data sets. In genomics, as repeated measures of phenotypes are crucial for understanding the onset of disease and its temporal pattern, longitudinal designs of omics data and phenotypes are being increasingly introduced. However, current statistical tests for longitudinal outcomes, especially for binary outcomes, depend heavily on the correct specification of the phenotype model. As many diseases are rare, efficient designs are commonly applied in epidemiological studies to recruit more cases. Despite the enhanced efficiency in the study sample, this non-random ascertainment sampling can be a major source of model misspecification that may lead to inflated type I error and/or power loss

in the association analysis. In transcriptomics, the analysis of single-cell RNA-seq data is facing its particular challenges due to low library size, high noise level, and prevalent dropout events. The purpose of this dissertation is to provide the methodological foundation to tackle the aforementioned challenges. We first propose a set of retrospective association tests for the identification of genetic loci associated with longitudinal binary traits. These tests are robust to different types of phenotype model misspecification and ascertainment sampling design which is common in longitudinal cohorts. We then extend these retrospective tests to variant-set tests for genetic rare variants that have low detection power by incorporating the variance component test and burden test into the retrospective test framework. Finally, we present a novel gene-graph based imputation method to impute dropout events in single-cell transcriptomic data to recover true gene expression level by borrowing information from adjacent genes in the gene graph.

Improving Risk Factor Identification of Human Complex Traits in Omics Data

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Weimiao Wu

Dissertation Director: Dr. Zuoheng Wang

June 2021

Contents

List of Figures

List of Tables

Acknowledgements

The past four and half years at Yale have been truly an extradentary journey and a rewarding experience for me. Even though the COVID pandemic kept us grounded for almost an entire year, it only makes me realized how beautiful our campus is and how much I missed it. Looking back, I would like to express my deepest gratitude to my thesis advisor, Dr. Zuoheng Wang, for her help, support, and encouragement during my PhD life. It is she who guided me through the training and taught me how to become a self-disciplined and independent researcher. Whenever I got stuck in projects or simply wanted to discuss a random idea, she was always available and willing to help. I always treasure those afternoons when we discussed the derivation of LSRAT statistics and their possible extensions. They were mind-blowing and really brought my understanding of region-based tests to a new level. Her rigorous attitude towards science and research had set a perfect example for me and will continue to influence my future career. I appreciate her broad range of knowledge and profound understanding of statistical theories which had enabled me to pursue various topics in genomics. I feel really lucky to have such a reliable mentor for my PhD study.

I would also like to express my sincere gratitude to Dr. Xiting Yan. I first met Dr. Yan at our joint journal club for single-cell RNA-seq data analysis. I was deeply fascinated by her extensive knowledge of transcriptomics data analysis and her passion for research. Later on, I was very excited to have the opportunity to work on several collaborative projects with her. Her kind, generous, and motivated personality had encouraged and comforted me during the tough times.

In addition, I would also like to thank my two other committee members, Dr. Hongyu Zhao, and Dr. Amy Justice, for their valuable insights and help with my thesis work. As the department chair, Dr. Zhao's dedication to science and research sets a role model for all students in this department. He has always been a source of motivation throughout my PhD training. Dr. Amy Justice is an expert for research in chronic disease and HIV infection related outcomes. She provided me with insightful comments on my research and generously provided VACS datasets for my projects. I would like to also thank all faculties and staff in the department of Biostatistics, especially Melanie S Elliot, the administrative director of Yale School of Public Health, for her kind and always timely help on my graduate student affairs.

Very special thanks go out to Dr. Rongling Wu and Dr. Yuehua Cui, without whose motivation and encouragement I would not have considered graduate training in biostatistics. It was under their guidance that I discovered my interests in statistical genetics when I was in college. They truly made a difference in my life.

Thanks also go out to my friends, whose companionship makes my life at Yale so colorful and joyful. Wenlan, your kitchen had not only warmed my stomach but also my heart. Yina, you are the sunshine that brought so much peace and warmth to my life. Wenxuan, the best study buddy, I still remember our overnight working on the course project in my apartment. Additionally, I would like to thank Yale Ballet instructor Ann Cowlin, who put efforts into continuing the dance classes virtually during the pandemic to keep us fit and boost our dopamine levels.

Finally and most importantly, I am grateful to my family for their unconditional support. For my twin sister, Weixiao, I always consider the luckiest thing that ever happened in my life is to have been born with a lifetime bestie. My sincerest regards to my parents and my grandparents for their love, sacrifices, and confidence in me. Lastly, I would like to thank my cat, Douhua, for burning all the midnight oil with me, with her feline grace and perhaps a negligible hint of impatience.

Dedicated to my beloved grandparents

even though they will most likely not read it

Chapter 1

Single-SNP Retrospective Association Tests For Longitudinal Binary Traits

1.1 Abstract

Longitudinal phenotypes have been increasingly available in genome-wide association studies (GWAS) and electronic health record-based studies for identification of genetic variants that influence complex traits over time. For longitudinal binary data, there remain significant challenges in gene mapping, including misspecification of the model for the phenotype distribution due to ascertainment. Here, we propose L-BRAT, a retrospective, generalized estimating equations-based method for genetic association analysis of longitudinal binary outcomes. We also develop RGMMAT, a retrospective, generalized linear mixed model-based association test. Both tests are retrospective score approaches in which genotypes are treated as random conditional on phenotype and covariates. They allow both static and time-varying covariates to be included in the analysis. Through simulations, we illustrated that retrospective association tests are robust to ascertainment and other types of phenotype model misspecification, and gain power over previous association methods. We applied L-BRAT and RGMMAT to a genome-wide association analysis of repeated measures of cocaine use in a longitudinal cohort. Pathway analysis implicated association with opioid signaling and axonal guidance signaling pathways. Lastly, we replicated important pathways in an independent

cocaine dependence case-control GWAS. Our results illustrate that L-BRAT is able to detect important loci and pathways in a genome scan and to provide insights into genetic architecture of cocaine use.

1.2 Introduction

Genome-wide association studies (GWAS) have successfully discovered many disease susceptibility loci and provided insights into the genetic architecture of numerous human complex diseases and traits. In some genetic epidemiological studies, longitudinally collected phenotype data are available. This is the case for many electronic health record (EHR)-based studies. As many of these studies continue to follow enrolled subjects (e.g. the UK Biobank (UKB) and the Million Veteran Program (MVP)), longitudinal phenotypes will be increasingly available with the passage of time, providing new data resources that require appropriate analytical tools for optimal analysis. Standard association tests that consider one time point or collapse repeated measurements into a single value such as an average do not capture the trajectory of phenotypic traits over time and may result in a loss of statistical power to detect genetic associations. In addition, the effects of time-varying covariates cannot be easily incorporated in such analyses. Recently, methodological developments for GWAS have proliferated to make full use of the available longitudinal data. For population cohorts, methods that account for dependence among observations from an individual include mixed effects models [1,2], generalized estimating equations (GEE) [3], growth mixture models[4,5], and empirical Bayes models [6]. Most of these approaches are prospective analyses and have been successfully applied to quantitative phenotypes.

As many diseases are rare, efficient designs, such as the case-control design, are commonly applied in epidemiological studies to recruit study subjects. Despite the enhanced efficiency in the study sample, non-random ascertainment can be a major source of model misspecification that may lead to inflated type I error and/or power loss in association analysis. The linear mixed model and the logistic mixed model do not perform well when the case-control ratio is unbalanced in large-scale genetic association studies [7]. Prospective analysis in which a population-based model is used ignores ascertainment bias and can result in compromised statistical inference. Furthermore, in the ascertained sample, the prospective approach conditional on the genotype and covariates may lose information when the joint distribution of the genotype and covariates carries additional information on whether the phenotype is associated with the genotype [8]. In this regard, several retrospective association methods have been proposed for analyzing ascertained population-based case-control studies [9,10], family-based studies of continuous traits [11], family-based case-control studies [12,13], and family-based longitudinal quantitative traits [14]. Compared to prospective tests, retrospective tests conditional on the phenotype and covariates are more robust to misspecification of the trait model[8].

To generalize case-control sampling, outcome-dependent sampling designs have become popular for binary data in longitudinal cohort studies [15–17]. However, association tests for longitudinally measured binary data are less well developed in GWAS. Here, we propose L-BRAT, a retrospective, GEE-based method for genetic association analysis of longitudinal binary outcomes. It requires specification of the mean of the outcome distribution and a working correlation matrix for repeated measurements.

L-BRAT is a retrospective score approach in which genotypes are treated as random conditional on the phenotype and covariates. Thus, it is robust to ascertainment and trait model misspecification. It allows both static and time-varying covariates to be included in the analysis. We note that GMMAT, a recently proposed prospective test using the logistic mixed model to control for population structure and cryptic relatedness in case-control studies [18], can be adapted for repeated binary data. For comparison, we also develop RGMMAT, a retrospective, generalized linear mixed model (GLMM)-based association test for longitudinal binary traits.

We performed simulation studies to evaluate the type I error and power of L-BRAT and RGMMAT, and compared them to the existing prospective methods. The results demonstrate that the retrospective association tests have better control of type I error when the phenotype model is misspecified, and are robust to various ascertainment schemes. Moreover, they are more powerful than the prospective tests. Finally, we applied L-BRAT and RGMMAT to a genome-wide association analysis of repeated measurements of cocaine use in a longitudinal cohort, the Veterans Aging Cohort Study (VACS), and replicated the results using data from an independent cocaine dependence case-control GWAS.

1.3 Materials and methods

Suppose a binary trait is measured over time on a study population of n individuals. We have their genome-wide measures of genetic variation. A set of covariates, static or dynamic, are also available. Let n_i be the number of repeated measures on individual i

and $N = \sum_{i=1}^{n} n_i$ be the total number of observations. For individual i, let $\boldsymbol{X}_{ij}$ and Y_{ij} be the p-dimensional covariate vector, assumed to include an intercept, and the binary response at time t_{ij}, respectively. In this setting, individuals are permitted to have measurements at different time points and different number of observations. We let $\boldsymbol{Y}$ denote the outcome vector of length N, and let $\boldsymbol{X}$ denote the $N \times p$ covariate matrix. Here, we focus on the problem of testing for association between a genetic variant and the longitudinal binary outcomes. Let $\boldsymbol{G}$ denote the vector of genotypes for the n individuals at the variant to be tested, where $G_i = 0$, 1, or 2 is the number of minor alleles of individual i at the variant.

1.3.1 Generalized estimating equations (GEE) model

We consider a GEE approach in which the mean of the outcome distribution, given the genotype and covariates, is specified as

$$E\left(Y_{ij} \mid \boldsymbol{G}, \boldsymbol{X}\right) = \mu_{ij}, \ \ \text{logit}\left(\mu_{ij}\right) = \boldsymbol{X}_{ij}^{T}\boldsymbol{\beta} + G_i\gamma, \ \ i = 1, \dots, n; \ j = 1, \dots, n_i, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of covariate effects and γ is a scalar parameter of interest representing the effect of the tested variant. Writing in a matrix form, we have the mean model

$$E(\boldsymbol{Y} \mid \boldsymbol{G}, \boldsymbol{X}) = \boldsymbol{\mu}, \ \ \text{logit}\,(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{G}\gamma, \tag{2}$$

where $\boldsymbol{B}$ is an $N \times n$ matrix representing the measurement clustering structure, and its (l, i)th entry B_{li} is an indicator of the lth entry of $\boldsymbol{Y}$ being a measurement on individual i. Here, the vector $\boldsymbol{B}\boldsymbol{G}$ is the vertically expanded genotype vector that maps the genotype data $\boldsymbol{G}$ from the individual level to the measurement level. The covariance structure of $\boldsymbol{Y}$ is given by

$$\mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{G}, \boldsymbol{X}) = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{1/2}, \tag{3}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}\{\mu_{1,1}(1-\mu_{1,1}), \ldots, \mu_{1,n_1}(1-\mu_{1,n_1}), \ldots, \mu_{n,1}(1-\mu_{n,1}), \ldots, \mu_{n,n_n}(1-\mu_{n,n_n})\}$ is an N-dimensional diagonal matrix and $\boldsymbol{\Sigma}$ is an $N \times N$ correlation matrix. The covariance specification in Eq. (3) ensures that the variance of the dichotomous response Y_{ij} depends on its mean in a way that is consistent with the Bernoulli distribution. To apply the GEE method, a working correlation structure such as independent, exchangeable, and first-order autoregressive (AR(1)) must be specified. For a given within-cluster correlation matrix $\boldsymbol{\Sigma}(\tau)$, which may depend on some parameter τ, the estimating equations for the unknown parameters $(\boldsymbol{\beta}, \gamma)$ are written as

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\beta}) \\ U(\gamma) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu}) \\ (\boldsymbol{BG})^T\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu}) \end{bmatrix}.$$

To detect association between the genetic variant and the phenotype, we consider a score approach to test $H_0: \gamma = 0$ against $H_1: \gamma \neq 0$. The null estimate of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}_0$, is the solution to a system of estimating equations $\boldsymbol{U}(\boldsymbol{\beta}) = 0$ under the constraint $\gamma = 0$, which can be computed iteratively between a Fisher scoring algorithm for $\boldsymbol{\beta}$ and the method of moments for τ until convergence. Then, the score function for γ is

$$U_0 = U(\gamma)_{|\widehat{\boldsymbol{\beta}}_0, 0, \hat{\tau}_0} = (\boldsymbol{BG})^T\hat{\boldsymbol{\Gamma}}_0^{1/2}\widehat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}}_0), \tag{4}$$

where $\widehat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_0$ are $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ evaluated at $(\boldsymbol{\beta}, \gamma, \tau) = (\widehat{\boldsymbol{\beta}}_0, 0, \hat{\tau}_0)$.

In the GEE approach, the prospective score statistic for testing $H_0: \gamma = 0$ takes the form

$$T_{GEE} = \frac{U_0^2}{\mathrm{Var}_0(U_0 \mid \boldsymbol{G}, \boldsymbol{X})} = \frac{\left[(\boldsymbol{BG})^T\hat{\boldsymbol{\Gamma}}_0^{1/2}\widehat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}}_0)\right]^2}{(\boldsymbol{BG})^T\boldsymbol{QBG}}, \tag{5}$$

where the null variance of U_0 is evaluated using a robust sandwich variance estimator, conditional on the genotype and covariates. Here $\boldsymbol{Q} = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}$, where $\boldsymbol{V} = \hat{\boldsymbol{\Gamma}}_0^{1/2}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}\widehat{\mathrm{Cov}}(\boldsymbol{Y})\hat{\boldsymbol{\Gamma}}_0^{-1/2}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{1/2}$ and the sample covariance of $\boldsymbol{Y}$, $\widehat{\mathrm{Cov}}(\boldsymbol{Y})$, is estimated by $(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_0)(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_0)^T$. Under the null hypothesis, the T_{GEE} test statistic has an asymptotic χ_1^2 distribution.

1.3.2 L-BRAT retrospective test

In what follows, we introduce a new GEE-based association testing method, L-BRAT (Longitudinal Binary-trait Retrospective Association Test). The L-BRAT test statistic is also based on the score function U_0 in Eq. (4). In contrast to the prospective GEE score test, L-BRAT takes a retrospective approach in which the variance of U_0 is assessed using a retrospective model of the genotype given the phenotype and covariates. An advantage of the retrospective approach is that the analysis is less dependent on the correct specification of the phenotype model. We assume that under the null hypothesis of no association between the genetic variant and the phenotype, the quasi-likelihood model of $\boldsymbol{G}$ conditional on $\boldsymbol{Y}$ and $\boldsymbol{X}$ is

$$E_0(\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}) = 2p\mathbf{1}_n, \ \ \mathrm{Var}_0(\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}) = \sigma_g^2\boldsymbol{\Phi}, \tag{6}$$

where p is the minor allele frequency (MAF) of the tested variant, $\mathbf{1}_n$ is a vector of length n with every element equal to 1, σ_g^2 is an unknown variance parameter, and $\boldsymbol{\Phi}$ is an $n \times n$ genetic relationship matrix (GRM) representing the overall genetic similarity between individuals due to population structure. Because $\boldsymbol{B}\mathbf{1}_n = \mathbf{1}_N$, which is the first column of $\boldsymbol{X}$ that encodes an intercept, and $\hat{\boldsymbol{\Gamma}}_0^{1/2}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_0)$, the N-dimensional

vector of transformed null phenotypic residuals, is orthogonal to the column space of $\boldsymbol{X}$, then the null mean model of $\boldsymbol{G}$ in Eq. (6) ensures that

$$E_0(U_0 \mid \boldsymbol{Y}, \boldsymbol{X}) = E_0(\boldsymbol{A}^T\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}) = 2p\boldsymbol{A}^T\mathbf{1}_n = 0,$$

where $\boldsymbol{A} = \boldsymbol{B}^T\hat{\boldsymbol{\Gamma}}_0^{1/2}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_0)$ is the individual-level transformed phenotypic residual vector of length n.

In model (6), the GRM $\boldsymbol{\Phi}$ can be obtained using genome-wide data, given by

$$\boldsymbol{\Phi} = \frac{1}{K}\sum_{k=1}^{K}\frac{\left(\boldsymbol{G}_{(k)} - 2\hat{p}_k\right)\left(\boldsymbol{G}_{(k)} - 2\hat{p}_k\right)^T}{2\hat{p}_k(1-\hat{p}_k)},$$

where K is the total number of genotyped variants, $\boldsymbol{G}_{(k)}$ is the genotype vector at the kth variant, and $\hat{p}_k$ is the estimated MAF, for example, $\hat{p}_k = \bar{G}_k/2$, the sample MAF at the kth variant. For the variant of interest, let $\hat{p} = \bar{G}/2$ be its sample MAF. Under Hardy-Weinberg equilibrium, the variance of the genotype is estimated by $\hat{\sigma}_g^2 = 2\hat{p}(1-\hat{p})$. Or we can use a more robust variance estimator (Jakobsdottir and McPeek 2013) given by

$$\tilde{\sigma}_g^2 = (n-1)^{-1}\boldsymbol{G}^T\boldsymbol{W}\boldsymbol{G},$$

where $\boldsymbol{W} = \boldsymbol{\Phi}^{-1} - \boldsymbol{\Phi}^{-1}\mathbf{1}_n(\mathbf{1}_n^T\boldsymbol{\Phi}^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n^T\boldsymbol{\Phi}^{-1}$. Finally, the L-BRAT test statistic can be defined as

$$\text{L-BRAT} = \frac{U_0^2}{\text{Var}_0(U_0 \mid \boldsymbol{Y}, \boldsymbol{X})} = \frac{\left(\boldsymbol{A}^T\boldsymbol{G}\right)^2}{\text{Var}_0\left(\boldsymbol{A}^T\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}\right)} = \frac{\left(\boldsymbol{A}^T\boldsymbol{G}\right)^2}{\hat{\sigma}_g^2\boldsymbol{A}^T\boldsymbol{\Phi}\mathbf{A}}. \quad (7)$$

Under regularity conditions, L-BRAT asymptotically follows a χ_1^2 distribution under the null hypothesis.

1.3.3 Generalized linear mixed model (GLMM)

The Generalized linear Mixed Model Association Test (GMMAT) was originally designed to use random effects in logistic mixed models to account for population

structure and cryptic relatedness in case-control studies[18].To extend the GMMAT method for case-control analysis to repeated binary data, we consider the following logistic mixed model:

$$\text{logit}\left(\mu_{ij}\right) = \boldsymbol{X}_{ij}^{T}\boldsymbol{\beta} + G_i\gamma + a_i + r_{ij}, \;\; i = 1, \dots, n; \; j = 1, \dots, n_i, \tag{8}$$

where $\mu_{ij} = P\left(Y_{ij} = 1 \,\middle|\, G_i, \boldsymbol{X}_{ij}, a_i, r_{ij}\right)$ is the probability of a binary response at time t_{ij} for individual i, conditional on his/her genotype, covariates, and random effects a_i and r_{ij}, $\boldsymbol{\beta}$ and γ are the same as defined in model (1), a_i is the individual random effect, and r_{ij} is the individual-specific time-dependent random effect. The two random effects were used to capture the correlation among repeated measures in gene-based association test for longitudinal traits [19]. Here, a_i's are assumed to be independent and $a_i \sim N(0, \sigma_a^2)$. The vector of time-dependent random effects $\boldsymbol{r}_i = (r_{i1}, \dots, r_{i,n_i})$ has a multivariate normal distribution, $\boldsymbol{r}_i \sim MVN(\boldsymbol{0}, \sigma_r^2\boldsymbol{R}_i)$, where an AR(1) structure is assumed for the correlation matrix $\boldsymbol{R}_i$, in which τ is the unknown parameter. The binary responses Y_{ij} are assumed to be independent given the random effects a_i and r_{ij}. Note that the first relatedness matrix of the random effects in the original GMMAT paper is genetic relationship matrix, but in our model for the longitudinal data, the two relatedness matrices correspond to the individual random effect and the individual specific time-dependent random effect.

To construct a score test for the null hypothesis $H_0: \gamma = 0$ vs. the alternative $H_1: \gamma \neq 0$, we use the penalized quasi-likelihood method [20] to fit the null logistic mixed model and obtain the null estimates of $\boldsymbol{\beta}, \sigma_a^2, \sigma_r^2$ and τ, denoted by $\widehat{\boldsymbol{\beta}}_0, \hat{\sigma}_a^2, \hat{\sigma}_r^2$ and $\hat{\tau}_0$ [18].

Similarly, the best linear unbiased predictor (BLUP) of random effects, $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{r}}$, can be obtained. Then, the resulting score function for γ is

$$S_0 = S(\gamma)_{|\widehat{\boldsymbol{\beta}}_0, 0, \hat{\sigma}_a^2, \hat{\sigma}_r^2, \hat{\tau}_0, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{r}}} = (\boldsymbol{BG})^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0), \tag{9}$$

where $\widehat{\boldsymbol{\mu}}_0 = \text{logit}^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_0 + \boldsymbol{B}\widehat{\boldsymbol{a}} + \widehat{\boldsymbol{r}})$ is a vector of fitted values under H_0.

In GMMAT, the null variance of the score S_0 is evaluated prospectively [18], i.e., $\text{Var}_0(S_0 \mid \boldsymbol{G}, \boldsymbol{X}) = (\boldsymbol{BG})^T\boldsymbol{PBG}$, where $\boldsymbol{P} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}$, and $\boldsymbol{\Psi} = \widehat{\boldsymbol{\Gamma}}_0^{-1} + \hat{\sigma}_a^2\boldsymbol{BB}^T + \hat{\sigma}_r^2\widehat{\boldsymbol{R}}$. Here $\widehat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{R}}$ are $\boldsymbol{\Gamma}$ and $\boldsymbol{R}$ evaluated at $(\boldsymbol{\beta}, \gamma, \sigma_a^2, \sigma_r^2, \tau) = \left(\widehat{\boldsymbol{\beta}}_0, 0, \hat{\sigma}_a^2, \hat{\sigma}_r^2, \hat{\tau}_0\right)$, where $\boldsymbol{\Gamma}$ is the same as defined in Eq. (3) and $\boldsymbol{R} = \text{diag}\{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n\}$ is a block diagonal matrix. The GMMAT test statistic can be written as

$$T_{GMMAT} = \frac{S_0^2}{\text{Var}_0(S_0 \mid \boldsymbol{G}, \boldsymbol{X})} = \frac{\left[(\boldsymbol{BG})^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)\right]^2}{(\boldsymbol{BG})^T\boldsymbol{PBG}}. \tag{10}$$

1.3.4 RGMMAT retrospective test

Like L-BRAT, we can construct a retrospective test to assess the significance of the GLMM score function of Eq. (9), which we call RGMMAT, based on the quasi-likelihood model of $\boldsymbol{G}$ in Eq. (6). Thus, we define the RGMMAT statistic by

$$\text{RGMMAT} = \frac{S_0^2}{\text{Var}_0(S_0 \mid \boldsymbol{Y}, \boldsymbol{X})} = \frac{(\boldsymbol{C}^T\boldsymbol{G})^2}{\text{Var}_0\left(\boldsymbol{C}^T\boldsymbol{G} \mid \boldsymbol{Y}, \boldsymbol{X}\right)} = \frac{(\boldsymbol{C}^T\boldsymbol{G})^2}{\hat{\sigma}_g^2\boldsymbol{C}^T\boldsymbol{\Phi}\boldsymbol{C}}, \tag{11}$$

where $\boldsymbol{C} = \boldsymbol{B}^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)$ is the n-dimensional vector of phenotypic residuals at the individual level by summing over all time points for an individual, and the phenotypic residuals are obtained by fitting the null logistic mixed model. Both the GMMAT and RGMMAT test statistics are assumed to have χ_1^2 asymptotic null distributions.

1.3.5 Simulation studies

We performed simulation studies to evaluate the type I error and power of the two retrospective tests we propose, and compared them to the prospective GEE and GMMAT methods. We also assessed sensitivity of L-BRAT and RGMMAT in the presence of model misspecification and ascertainment. In the simulations, we considered two different trait models and three different ascertainment schemes. Because both the L-BRAT and GEE methods require specification of a working correlation matrix, we implemented three working correlation structures: (1) independent, (2) AR(1), and (3) a mixture of exchangeable and AR(1).

To generate genotypes, we first simulated 10,000 chromosomes over a 1 Mb region using a coalescent model that mimics the linkage disequilibrium (LD) and recombination rates of the European population [21]. We then randomly selected 1,000 non-causal single nucleotide polymorphisms (SNPs) with MAF > 0.05. In addition, we generated two causal SNPs that were assumed to influence the trait value with epistasis. In the type I error simulations, we tested association at the 1,000 non-causal SNPs. In each simulation setting, we generated 1,000 sets of phenotypes at five time points. Putting together, 10^6 replicates were used for the type I error evaluation. In the power simulations, we tested the first of the two causal SNPs and empirical power was assessed using 1,000 simulation replicates. In all tests considered, the genotypes at the untested causal SNP(s) were assumed to be unobserved.

Trait models

We simulated two types of binary trait models at five time points, in which the two unlinked causal SNPs were assumed to act on the phenotype epistatically. The first type is a logistic mixed model, given by

$$Y_{ij}|\boldsymbol{X}_{ij}, G_{i(1)}, G_{i(2)}, a_i, r_{ij} \sim \text{Bernoulli}\left(\mu_{ij}\right),$$

$$\text{logit}\left(\mu_{ij}\right) = -2.5 + 0.2(j-1) + 0.5X_{ij(1)} + 0.5X_{i(2)} + \theta 1_{\{G_{i(1)}>0, G_{i(2)}>0\}} + a_i + r_{ij},$$

where $X_{ij(1)}$ is a continuous, time-varying covariate generated independently from a standard normal distribution, $X_{i(2)}$ is a binary, time-invariant covariate taking values 0 or 1 with a probability of 0.5, $G_{i(1)}$ and $G_{i(2)}$ are the genotypes of individual i at the two causal SNPs, θ is a scalar parameter encoding the effect of the causal SNPs, $1_{\{G_{i(1)}>0, G_{i(2)}>0\}}$ is an indicator function that takes value 1 when individual i has at least one copy of the minor allele at both the causal SNPs, a_i and r_{ij} are the individual-level time-independent and time-dependent random effects, respectively. Here we assume $a_i \sim N(0,\ \sigma_a^2)$ and $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{i5}) \sim MVN(\boldsymbol{0}, \sigma_r^2 \boldsymbol{R})$, where $\boldsymbol{R}$ is a 5×5 correlation matrix specified by the AR(1) structure with a correlation coefficient τ. The two causal SNPs are assumed to be unlinked with MAFs 0.1 and 0.5, respectively. The variance components are set to $\sigma_a^2 = \sigma_r^2 = 0.64$ and $\tau = 0.7$.

The second type of trait model we considered is a liability threshold model in which an underlying continuous liability determines the outcome value based on a threshold. Specifically, the phenotype Y_{ij} is given by

$$Y_{ij} = 1 \text{ if } L_{ij} > 0,$$

$$\text{with } L_{ij} = -2.0 + 0.2(j-1) + 0.5X_{ij(1)} + 0.5X_{i(2)} + \theta 1_{\{G_{i(1)}>0, G_{i(2)}>0\}} + a_i + r_{ij} + e_{ij},$$

where L_{ij} is the underlying liability for individual i at time t_{ij}, and $e_{ij} \sim N(0, \sigma_e^2)$ represents independent noise, with $\sigma_e^2 = 0.64$. All other parameters are the same as those in the logistic mixed model.

In both models, we included a time effect and assumed that the mean of the outcome increases with time. The effect of the causal SNPs was set to $\theta = 0.34$ in the type I error simulations. For the power evaluation, we considered a range of values for θ, where we set $\theta = 0.3$, 0.32, 0.34, 0.36, and 0.38. At the given parameter values, the prevalence of the event of interest ranges from 12.8% to 27.7% over time. The proportion of the phenotypic variance explained by the two causal SNPs ranges from 0.69% to 1.10% in the logistic mixed model, and from 0.49% to 0.78% in the liability threshold model.

Sampling designs

We considered three different sampling designs. In the "random" sampling scheme, the sample contains 2,000 individuals that were randomly selected from the population regardless of their phenotypes. Thus, ascertainment is population based. In the "baseline" sampling scheme, we sampled 1,000 case subjects and 1,000 control subjects according to their outcome value at baseline only. In the "sum" sampling scheme, individuals were stratified into three strata (1, 2, and 3) based on a total count that sums each subject's response over time, where samples in stratum 1 never experienced the event of interest, i.e., $\sum_j Y_{ij} = 0$, samples in stratum 2 sometimes experienced the event, i.e., $0 < \sum_j Y_{ij} < n_i$, and samples in stratum 3 always experienced the event, i.e., $\sum_j Y_{ij} = n_i$. Following the outcome-dependent sampling design for longitudinal binary data [17], we selected

100, 1,800, and 100 individuals from the three strata respectively to oversample subjects who have response variation over the course of the study.

1.3.6 Application to cocaine use data from VACS

We illustrated the utility of our proposed methods by analyzing a GWAS dataset of cocaine use from VACS [22]. VACS is a multi-center, longitudinal observational study of HIV infected and uninfected veterans whose primary objective is to understand the risk of alcohol and other substance abuse in individuals with HIV infection. We analyzed longitudinal cocaine use in patient surveys collected at six clinic visits on 2,470 participants. Among them, 69.8% are African Americans (AAs), 19.3% are European Americans (EAs), and 10.9% are of other races. We considered the responses at each visit as 0 if individuals had never tried cocaine or had not used cocaine in the last year, and as 1 if individuals had used cocaine in the last year. The proportion of case subjects at each visit ranges from 13.7% ($n = 192$) to 24.3% ($n = 526$), and the missing rate at each visit ranges from 3.0% to 44.2%.

All samples were genotyped on the Illumina OmniExpress BeadChip. After data cleaning, there are 2,458 individuals available for genotype imputation. IMPUTE2 [23] was used for imputation using the 1000 Genomes Phase 3 data as a reference panel. We excluded subjects who did not meet either of the following criteria: (1) completeness (i.e., proportion of successfully imputed SNPs) > 95% and (2) empirical self-kinship < 0.525 (i.e., empirical inbreeding coefficient < 0.05). Based on the above criteria, 2,231 individuals were retained in the analysis, with 2,114 males and 117 females, of whom 1,557 are AAs, 431 are EAs, and 243 are of other races. There are 1,433 individuals who had never used cocaine during the study period, 639 individuals who sometimes used

cocaine, i.e., exhibited response variation, and 159 individuals who had used cocaine at least once every year over the course of the study. SNPs that satisfied all of the following quality-control conditions were included in the analysis: (1) call rate > 95%, (2) Hardy-Weinberg χ^2 statistic P-value > 10^{-6}, and (3) MAF > 1%. All together there are a final set of 10,215,072 SNPs retained in the analysis. The VACS dataset, both the genotype file in the plink format and the phenotype files including the longitudinally measured cocaine use and the covariates, will be deposited to dbGap (https://www.ncbi.nlm.nih.gov/gap).

Pathway and enrichment analyses

We then performed pathway analysis on the top SNPs for which at least one of the longitudinal tests had a P-value $< 5 \times 10^{-5}$ using the Ingenuity Pathway Analysis (IPA). The Ingenuity database gathers information from manually reviewed literature, as well as large public databases. In this analysis, the top SNPs' RSID were uploaded into the IPA and mapped, if possible, to the reference set in the Ingenuity knowledge. The IPA performs a Fisher's Exact test to determine whether the submitted SNP list belongs to genes of a particular function annotation more than expected by chance. We report below both Fisher's exact test P-value and adjusted P-value using Benjamini-Hochberg method for multiple testing adjusting for the number of ontologies tested. We consider pathways with adjusted P-value less than 0.05 to be significant. We also performed an enrichment analysis to see whether the top SNPs in our analysis are more likely to regulate brain gene expression.

Replication data

We used an independent cocaine dependence case-control GWAS from the Yale-Penn study [24] to replicate the top findings from our longitudinal analysis results in

VACS. The summary statistics from the Yale-Penn cocaine dependence GWAS were obtained. Note that the lifetime cocaine dependence diagnosis was made using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [25], which is different from the outcome used in VACS, and there were no longitudinal phenotype measures in Yale-Penn. Pathway analysis using IPA was applied to the summary statistics of Yale-Penn on the top SNP list identified from VACS. The Fisher's exact test P-values were calculated for each pathway to evaluate if there were more associated SNPs than would be expected by chance.

1.4 Results

1.4.1 Type I error assessment

To assess type I error, we tested for association at unlinked and unassociated SNPs. Table 1.1 gives the empirical type I error of the L-BRAT, RGMMAT, GEE, and GMMAT tests, based on 10^6 replicates, at the nominal type I error level α, for $\alpha = 0.05$, 0.01, 0.001, and 0.0001. In all simulations, the type I error of the two retrospective tests, L-BRAT and RGMMAT, exhibited no inflation at any of the nominal levels considered. In contrast, the prospective GEE tests, regardless of the choice of working correlation, had inflated type I error at most of the nominal levels in all settings. This is likely due to the fact that the asymptotic distribution of robust sandwich variance estimators used in GEE are not well calibrated. The inflated type I error was also reported in longitudinal GWAS with quantitative traits using GEE [3]. In GMMAT, the type I error was much lower than the nominal level when $\alpha = 0.05$, 0.01, 0.001, and 0.0001. These results demonstrate that the two retrospective tests, L-BRAT and RGMMAT, are robust to trait

model misspecification and ascertainment, whereas GEE has type I error inflation and GMMAT is overly conservative. Overall, the choice of the working correlation matrix does not have much impact on the type I error of the L-BRAT method.

Table 1.1. Empirical type I error of L-BRAT, RGMMAT, GEE, and GMMAT, based on 10^6 replicates.

Test	Level	Logistic Mixed Model			Liability Threshold Model		
		Random	Baseline	Sum	Random	Baseline	Sum
GEE (ind)	0.05	**5.38×10^{-2}**	**5.08×10^{-2}**	**5.27×10^{-2}**	**5.36×10^{-2}**	**5.19×10^{-2}**	**5.38×10^{-2}**
	0.01	**1.18×10^{-2}**	**1.04×10^{-2}**	**1.13×10^{-2}**	**1.17×10^{-2}**	**1.07×10^{-2}**	**1.17×10^{-2}**
	0.001	**1.32×10^{-3}**	**1.16×10^{-3}**	**1.23×10^{-3}**	**1.37×10^{-3}**	**1.14×10^{-3}**	**1.37×10^{-3}**
	0.0001	**1.67×10^{-4}**	**1.28×10^{-4}**	**1.43×10^{-4}**	**1.34×10^{-4}**	**1.36×10^{-4}**	**1.76×10^{-4}**
GEE (AR1)	0.05	**5.36×10^{-2}**	5.02×10^{-2}	**5.26×10^{-2}**	**5.34×10^{-2}**	**5.17×10^{-2}**	**5.37×10^{-2}**
	0.01	**1.16×10^{-2}**	**1.04×10^{-2}**	**1.12×10^{-2}**	**1.16×10^{-2}**	**1.06×10^{-2}**	**1.17×10^{-2}**
	0.001	**1.31×10^{-3}**	**1.13×10^{-3}**	**1.21×10^{-3}**	**1.36×10^{-3}**	**1.14×10^{-3}**	**1.36×10^{-3}**
	0.0001	**1.73×10^{-4}**	1.19×10^{-4}	**1.37×10^{-4}**	**1.32×10^{-4}**	**1.35×10^{-4}**	**1.78×10^{-4}**
GEE(mix)	0.05	**5.34×10^{-2}**	**5.07×10^{-2}**	**5.26×10^{-2}**	**5.34×10^{-2}**	**5.19×10^{-2}**	**5.37×10^{-2}**
	0.01	**1.17×10^{-2}**	**1.04×10^{-2}**	**1.13×10^{-2}**	**1.16×10^{-2}**	**1.07×10^{-2}**	**1.17×10^{-2}**
	0.001	**1.29×10^{-3}**	**1.17×10^{-3}**	**1.22×10^{-3}**	**1.38×10^{-3}**	**1.14×10^{-3}**	**1.36×10^{-3}**
	0.0001	**1.70×10^{-4}**	**1.29×10^{-4}**	**1.37×10^{-4}**	**1.31×10^{-4}**	**1.30×10^{-4}**	**1.70×10^{-4}**
GMMAT	0.05	3.89×10^{-2}	3.53×10^{-2}	4.76×10^{-2}	4.80×10^{-2}	4.89×10^{-2}	4.91×10^{-2}
	0.01	6.07×10^{-3}	5.24×10^{-3}	9.08×10^{-3}	9.29×10^{-3}	9.51×10^{-3}	9.33×10^{-3}
	0.001	4.29×10^{-4}	3.74×10^{-4}	7.84×10^{-4}	8.63×10^{-4}	8.96×10^{-4}	8.33×10^{-4}
	0.0001	2.20×10^{-5}	2.20×10^{-5}	6.80×10^{-5}	6.30×10^{-5}	9.10×10^{-5}	8.80×10^{-5}
L-BRAT (ind)	0.05	4.93×10^{-2}	4.91×10^{-2}	4.98×10^{-2}	5.01×10^{-2}	4.99×10^{-2}	4.98×10^{-2}
	0.01	9.45×10^{-3}	9.60×10^{-3}	9.84×10^{-3}	9.90×10^{-3}	9.75×10^{-3}	9.55×10^{-3}
	0.001	8.30×10^{-4}	9.78×10^{-4}	9.24×10^{-4}	9.55×10^{-4}	9.45×10^{-4}	8.78×10^{-4}
	0.0001	7.20×10^{-5}	9.50×10^{-5}	8.20×10^{-5}	8.20×10^{-5}	9.40×10^{-5}	9.20×10^{-5}
L-BRAT	0.05	4.93×10^{-2}	4.88×10^{-2}	4.97×10^{-2}	4.99×10^{-2}	4.98×10^{-2}	4.97×10^{-2}

(AR1)	0.01	9.48×10^{-3}	9.72×10^{-3}	9.78×10^{-3}	9.84×10^{-3}	9.76×10^{-3}	9.55×10^{-3}
	0.001	8.26×10^{-4}	9.62×10^{-4}	9.22×10^{-4}	9.17×10^{-4}	9.47×10^{-4}	8.48×10^{-4}
	0.0001	8.80×10^{-5}	9.60×10^{-5}	8.20×10^{-5}	7.10×10^{-5}	1.02×10^{-4}	8.90×10^{-5}
L-BRAT (mix)	0.05	4.93×10^{-2}	4.91×10^{-2}	4.99×10^{-2}	5.01×10^{-2}	4.98×10^{-2}	4.98×10^{-2}
	0.01	9.57×10^{-3}	9.61×10^{-3}	9.86×10^{-3}	9.88×10^{-3}	9.79×10^{-3}	9.54×10^{-3}
	0.001	8.35×10^{-4}	9.86×10^{-4}	9.26×10^{-4}	9.57×10^{-4}	9.37×10^{-4}	8.78×10^{-4}
	0.0001	8.20×10^{-5}	1.01×10^{-4}	8.60×10^{-5}	7.40×10^{-5}	9.70×10^{-5}	8.90×10^{-5}
RGMMAT	0.05	4.72×10^{-2}	4.91×10^{-2}	4.98×10^{-2}	4.93×10^{-2}	4.99×10^{-2}	4.98×10^{-2}
	0.01	8.76×10^{-3}	9.64×10^{-3}	9.85×10^{-3}	9.63×10^{-3}	9.78×10^{-3}	9.55×10^{-3}
	0.001	7.20×10^{-4}	9.52×10^{-4}	9.09×10^{-4}	9.12×10^{-4}	9.43×10^{-4}	8.75×10^{-4}
	0.0001	6.80×10^{-5}	8.90×10^{-5}	8.20×10^{-5}	7.70×10^{-5}	9.10×10^{-5}	9.30×10^{-5}

1.4.2 Power comparison

To compare power, we considered five effect sizes at the two causal SNPs, and tested association between the trait and the first causal SNP. Empirical power was calculated at the significance level 10^{-3}, based on 1,000 simulated replicates. Figure 1 demonstrates the power results for each method. In all the simulation settings, the retrospective tests consistently had higher power than the prospective tests. The L-BRAT association tests under three different working correlation structures had similar power. The RGMMAT method also achieved high power. In contrast, the prospective GEE methods had the lowest power in all settings except under the baseline sampling and the liability threshold model, in which GMMAT performed the worst in power. Overall, we found that the baseline sampling scheme generated the highest power under different trait models, while the sum sampling scheme had a power gain over the random sampling scheme under the logistic mixed model, but was less powerful under the liability threshold model. These

results suggest that L-BRAT and RGMMAT outperform the prospective tests, and the power of L-BRAT is not sensitive to the choice of the working correlation structure.

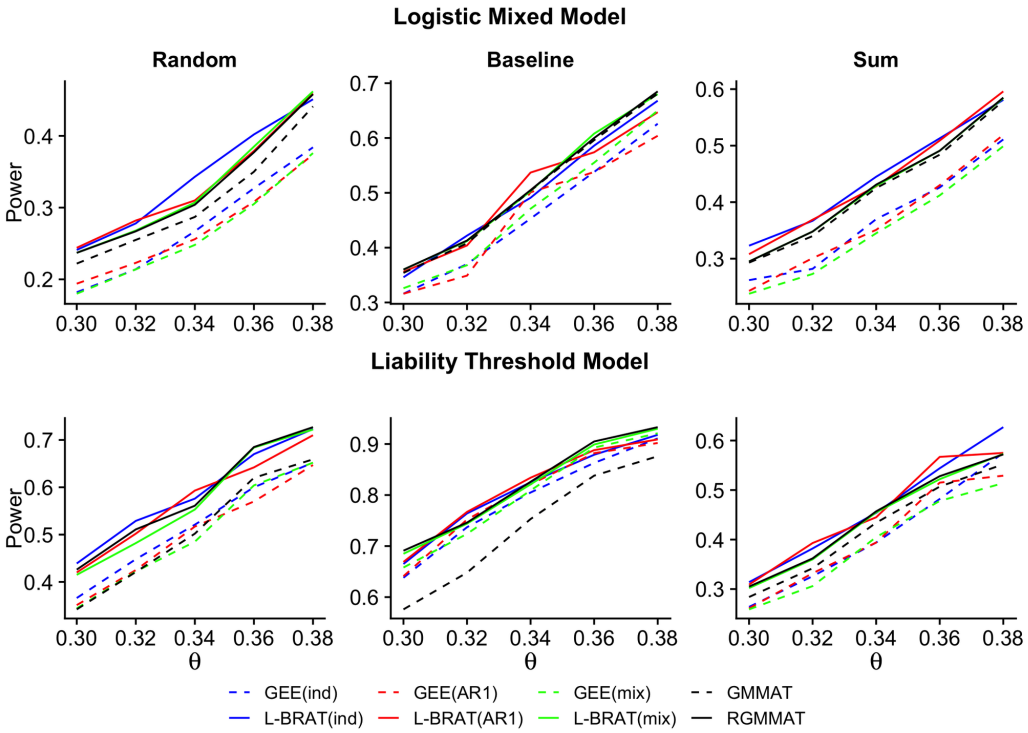


Figure 1.1. Empirical power of L-BRAT, RGMMAT, GEE, and GMMAT. Power is based on 1,000 simulated replicates at five time points with $\alpha = 10^{-3}$. In the upper panel, the trait is simulated by the logistic mixed model, and in the lower panel, it is by the liability threshold model. Power results are demonstrated in samples of 2,000 individuals according to three different ascertainment schemes: random, baseline, and sum. This figure appears in color in the electronic version of this article.

1.4.3 Analysis of Longitudinal Cocaine Use Data from VACS

Genome-wide association testing for longitudinal cocaine use was performed using L-BRAT, RGMMAT, and the prospective GEE and GMMAT tests in the entire VACS sample. Sex, age at baseline, HIV status, and time were included as covariates in the analysis. The top ten principal components (PCs) that explained 89.4% of the total genetic variation were included as covariates to control for population structure. We

considered two working correlation structures: independent and AR(1). For the L-BRAT and RGMMAT methods, the GRM was calculated using the LD pruned SNPs with MAF > 0.05.

To compare the performance of longitudinal association analysis with that of univariate analysis on the summary metrics of cocaine use in VACS, we considered two alternative cocaine phenotypes: baseline and trajectories. Longitudinal cocaine use trajectories were obtained using a growth mixture model that clusters longitudinal data into discrete growth trajectory curves [26]. We fit a logistic model with a polynomial function of time. The number of groups was chosen based on the Bayesian information criterion (BIC). Each individual was then assigned to the trajectory with the highest probability of membership. Figure 2 shows the four cocaine use trajectory groups identified in the VACS sample. They were labeled as mostly never (0, $n = 1{,}682$), moderate decrease (1, $n = 296$), elevated chronic (2, $n = 86$), and mostly frequent (3, $n = 167$). We used CARAT, a case-control retrospective association test [10], for the analysis of cocaine use at baseline, adjusted for sex, age at baseline, and HIV status. Cumulative logit model was used to test for association between the four ordered cocaine use trajectory groups and each of the SNPs, with adjustment for sex, age at baseline, HIV status, and the top ten PCs.

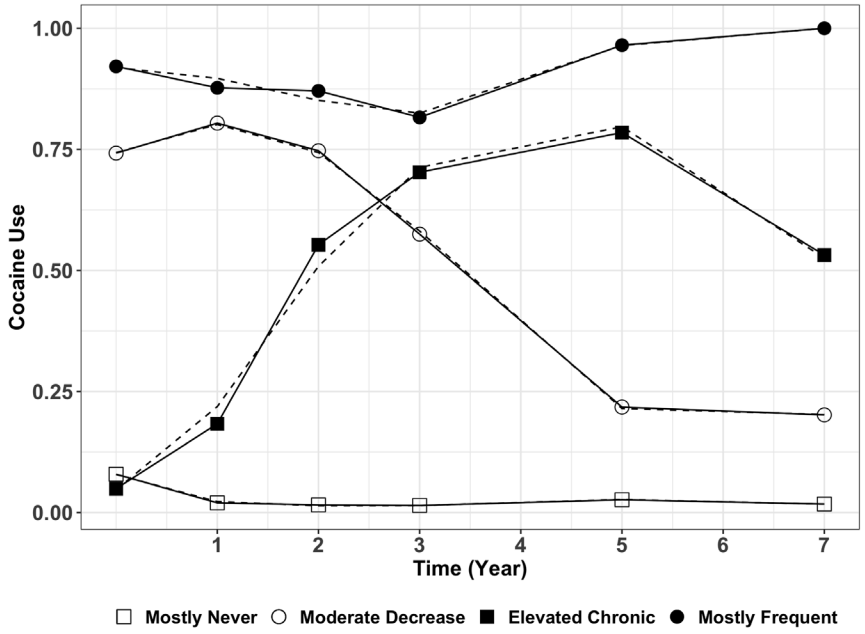


Figure 1.2. Group-based cocaine use trajectories in VACS. Dashed lines represent the estimated trajectories, solid lines represent the observed mean cocaine use for each trajectory group. Time is the number of years since the baseline visit.

None of the retrospective tests exhibited evidence of inflation in the quantile-quantile (Q-Q) plot. The genomic control inflation factors were 0.993 and 0.991 for the L-BRAT genome scan under the independent and AR(1) working correlation, respectively, and 0.984 for the RGMMAT analysis. The prospective GEE tests showed some evidence of deflation in the Q-Q plot. The genomic control factors were 0.938 and 0.937 for the GEE tests under the independent and AR(1) working correlation. The most conservative test was GMMAT, with a genomic control factor 0.802.

Table 1.2 reports the results for SNPs for which at least one of the longitudinal tests gives a P-value $< 2 \times 10^{-7}$. Among them, the L-BRAT tests produced the smallest P-values, RGMMAT and the trajectory-based analysis had comparable results, while GEE, GMMAT, and CARAT generated much larger P-values. Among the top SNPs listed in

Table 1.2, there are two SNPs, rs551879660 and rs150191017, located at 3p12 and 13q12 respectively, that reach the genome-wide significance ($P = 2.00 \times 10^{-8}$ and 3.77×10^{-8}, respectively). Each of these SNPs was reported to have MAF < 1% in the 1000 Genomes (MAF = 0.68% and 0.98%, respectively). The MAFs of the two SNPs were 1.2% and 1.1% in the entire VACS sample, respectively, and were slightly higher in the AA sample (MAF = 1.6% and 1.5%, respectively). Although both SNPs have MAF > 1%, given the small sample size of VACS, there is limited information on them. SNP rs150191017 is located 31.5 kb from the gene *AL161616.2* which was reported to be associated with venlafaxine treatment response in a generalized anxiety disorder GWAS [27]. A cluster of five SNPs in LD, rs76386683, rs114386843, rs186274502, rs376616438, and rs187855416, located at 9q33, showed association with longitudinal cocaine use ($P = 1.85 \times 10^{-7} - 1.93 \times 10^{-7}$). They are near *OR1L4*, an olfactory receptor gene that was reported to be associated with major depressive disorder [28]. A cluster of olfactory receptor genes between *OR3A1* and *OR3A2* that belong to the olfactory receptor gene family were identified in a recent GWAS of cocaine dependence and related traits [24]. The other three SNPs, rs188222191, rs1014278, rs75132056, are located at 5q21 ($P = 1.28 \times 10^{-7}$, 1.43×10^{-7} and 8.92×10^{-8}, respectively), close to the gene *EFNA5*, which was identified in several GWAS to be associated with bipolar disorder and schizophrenia [29]. There was also evidence of association with rs114629793 ($P = 8.65 \times 10^{-8}$). This SNP is in an intron of the gene encoding *PSD3*, located at 8p22. Recently, two schizophrenia GWAS have identified association between *PSD3* and schizophrenia [30,31], and one study has shown that *PSD3* is associated with paliperidone treatment response in schizophrenic patients [32]. Gene network analysis

revealed that *PSD3* is one of the differentially expressed hub genes that involve dysfunction of brain reward circuitry in cocaine use disorder [33].

Table 1.2. SNPs with P-value < 2×10−7 in at least one of the longitudinal tests in the entire VACS sample. The smallest P-value among all tests at the given SNPs are in bold. [a] CARAT applied to cocaine use at baseline, [b] Cumulative logit model applied to the four ordered cocaine use trajectory group.

Chr.	Gene Region	SNP	Position	MAF	GEE (ind)	GEE (AR1)	GMMAT	L-BRAT (ind)	L-BRAT (AR1)	RGMMAT	CARAT[a] (BL)	CL[b] (traj)
3	*NIPA2P2*	rs551879660	75,146,492	0.012	1.87×10^{-4}	7.14×10^{-4}	9.07×10^{-4}	2.00×10^{-8}	3.19×10^{-6}	4.13×10^{-5}	5.78×10^{-4}	3.35×10^{-5}
5	*EFNA5*	rs188222191	105,411,547	0.042	6.86×10^{-6}	1.65×10^{-5}	8.87×10^{-5}	1.28×10^{-7}	4.17×10^{-7}	2.69×10^{-6}	8.95×10^{-5}	2.72×10^{-5}
		rs1014278	105,471,506	0.057	1.02×10^{-5}	1.10×10^{-5}	1.24×10^{-4}	1.50×10^{-7}	1.43×10^{-7}	4.88×10^{-6}	5.94×10^{-5}	3.00×10^{-5}
		rs75132056	105,480,442	0.05	1.05×10^{-5}	2.42×10^{-5}	1.89×10^{-4}	8.92×10^{-8}	2.89×10^{-7}	8.55×10^{-6}	2.59×10^{-4}	2.31×10^{-5}
8	*PSD3*	rs114629793	18,403,754	0.012	3.12×10^{-4}	4.73×10^{-4}	1.44×10^{-4}	8.65×10^{-8}	3.60×10^{-7}	2.82×10^{-6}	5.12×10^{-4}	3.06×10^{-6}
9	*OR1L4*	rs76386683	125,467,023	0.012	1.48×10^{-4}	9.15×10^{-5}	2.86×10^{-4}	1.03×10^{-6}	1.93×10^{-7}	5.92×10^{-6}	4.80×10^{-4}	3.30×10^{-6}
		rs114386843	125,469,425	0.012	1.47×10^{-4}	9.05×10^{-5}	2.82×10^{-4}	1.01×10^{-6}	1.88×10^{-7}	5.78×10^{-6}	4.75×10^{-4}	3.22×10^{-6}
		rs186274502	125,471,416	0.012	1.47×10^{-4}	9.05×10^{-5}	2.82×10^{-4}	1.01×10^{-6}	1.88×10^{-7}	5.78×10^{-6}	4.75×10^{-4}	3.22×10^{-6}
		rs376616438	125,472,267	0.012	1.44×10^{-4}	8.95×10^{-5}	2.77×10^{-4}	9.79×10^{-7}	1.85×10^{-7}	5.62×10^{-6}	4.79×10^{-4}	3.20×10^{-6}
		rs187855416	125,474,459	0.012	1.44×10^{-4}	8.95×10^{-5}	2.77×10^{-4}	9.79×10^{-7}	1.85×10^{-7}	5.62×10^{-6}	4.79×10^{-4}	3.20×10^{-6}
11	*AP000851.1*	rs139780693	102,509,700	0.03	2.60×10^{-5}	1.04×10^{-5}	2.78×10^{-4}	5.83×10^{-7}	1.26×10^{-7}	1.35×10^{-5}	1.06×10^{-4}	2.00×10^{-6}
13	*AL161616.2*	rs150191017	31,962,649	0.011	4.26×10^{-5}	9.72×10^{-5}	7.32×10^{-5}	3.77×10^{-8}	3.09×10^{-7}	7.87×10^{-7}	3.74×10^{-4}	5.48×10^{-7}

We further analyzed the data separately in each population, adjusted for the top ten PCs obtained within the group, and then combined the results from the three groups by meta-analysis using the optimal weights for score statistics that have essentially the same power as the inverse variance weighting [34]. The results from the three groups (AAs,

EAs and other races) were combined by meta-analysis. The meta-analysis P-values were of the same order of magnitude as that obtained from the entire sample adjusted for population structure for each longitudinal test (Table 1.3). All the top twelve SNPs listed in Table 1.3 had a meta-analysis P-value $< 8 \times 10^{-7}$ in at least one of the longitudinal tests. Among them, the L-BRAT test with either an independent or AR(1) working correlation gave the smallest meta-analysis P-values.

Table 1.3. Meta-analysis results of the top twelve SNPs from Table 1.2 in the VACS data. The smallest P-value among all tests at the given SNPs are in bold.

Chr	Gene Region	SNP	Position	GEE (ind)	GEE (AR1)	GMMAT	L-BRAT (ind)	L-BRAT (AR1)	RGMMAT
3	*NIPA2P2*	rs551879660	75,146,492	1.81×10^{-4}	5.86×10^{-4}	8.98×10^{-4}	5.26×10^{-8}	6.41×10^{-6}	6.49×10^{-5}
5	*EFNA5*	rs188222191	105,411,547	7.57×10^{-6}	1.28×10^{-5}	1.80×10^{-4}	2.55×10^{-7}	5.52×10^{-7}	1.10×10^{-5}
		rs1014278	105,471,506	1.26×10^{-5}	8.44×10^{-6}	3.15×10^{-4}	1.03×10^{-6}	5.59×10^{-7}	2.44×10^{-5}
		rs75132056	105,480,442	1.31×10^{-5}	2.00×10^{-5}	4.24×10^{-4}	7.31×10^{-7}	1.27×10^{-6}	3.56×10^{-5}
8	*PSD3*	rs114629793	18,403,754	2.92×10^{-4}	4.31×10^{-4}	1.66×10^{-4}	1.79×10^{-7}	7.98×10^{-7}	6.83×10^{-6}
9	*OR1L4*	rs76386683	125,467,023	1.44×10^{-4}	8.78×10^{-5}	3.75×10^{-4}	2.32×10^{-6}	5.12×10^{-7}	1.46×10^{-5}
		rs114386843	125,469,425	1.42×10^{-4}	8.62×10^{-5}	3.68×10^{-4}	2.25×10^{-6}	4.97×10^{-7}	1.41×10^{-5}
		rs186274502	125,471,416	1.42×10^{-4}	8.62×10^{-5}	3.68×10^{-4}	2.25×10^{-6}	4.97×10^{-7}	1.41×10^{-5}
		rs376616438	125,472,267	1.39×10^{-4}	8.51×10^{-5}	3.60×10^{-4}	2.18×10^{-6}	4.86×10^{-7}	1.37×10^{-5}
		rs187855416	125,474,459	1.39×10^{-4}	8.51×10^{-5}	3.60×10^{-4}	2.18×10^{-6}	4.86×10^{-7}	1.37×10^{-5}
11	*AP000851.1*	rs139780693	102,509,700	1.15×10^{-5}	4.16×10^{-6}	1.07×10^{-4}	4.04×10^{-7}	6.05×10^{-8}	4.41×10^{-6}
13	*AL161616.2*	rs150191017	31,962,649	3.55×10^{-5}	6.77×10^{-5}	1.26×10^{-4}	6.68×10^{-8}	5.80×10^{-7}	3.12×10^{-6}

The smallest P-value among all tests at the given SNPs are in bold.

Pathway and enrichment analysis results

We identified two significant canonical pathways that belong to the neurotransmitters and nervous system signaling. The first one is the opioid signaling pathway ($P = 1.41 \times 10^{-4}$, adjusted P = 0.010), which plays an important role in opioid tolerance and dependence. Studies have shown that chronic administration of cocaine and opioids are associated with changes in dopamine transporters and opioid receptors in various brain regions [35,36] . The second significant pathway is the axonal guidance signaling pathway ($P = 2.54 \times 10^{-4}$, adjusted P = 0.012), which is critical for neural development. The mRNA expression levels of axon guidance molecules have been found to be altered in some brain regions of cocaine-treated rats, which may contribute to drug abuse-associated cognitive impairment [37,38]. Each of the two pathways remained significant when we performed pathway analysis, using the same P-value cutoff value to select top SNPs, based on the L-BRAT results generated under the independence and AR(1) working correlation, respectively. In contrast, only the opioid signaling pathway was significant based on the results from the GEE analysis using the independent working correlation, and only the axonal guidance signaling pathway was significant based on the RGMMAT results, whereas neither of them remained significant based on the GMMAT results and that from the GEE analysis with an AR(1) working correlation. These results demonstrate that L-BRAT provides more informative association results to help identify biological relevant pathways.

Lastly, we performed an enrichment analysis to see whether the top SNPs in our analysis are more likely to regulate brain gene expression. We considered the local expression quantitative trait loci (*cis*-eQTLs) reported in 13 human brain regions from the

Genotype-Tissue Expression (GTEx) project [39,40], including *amygdala, anterior cingulate cortex, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens, putamen, spinal cord*, and *substantia nigra*. Fisher's exact test was used to assess the enrichment of eQTLs (FDR < 0.05) in the top 2,778 SNPs for which at least one of the longitudinal tests had a P-value $< 10^{-4}$ in the VACS sample. Among the 13 brain regions, *amygdala* is the only region in which eQTLs showed significant enrichment in our top SNP list (odds ratio = 2.06, $P = 3.0 \times 10^{-5}$).

Replication of top findings

Nevertheless, we performed pathway analysis using the SNP summary statistics of Yale-Penn to replicate the two pathways identified in the VACS sample. Among the top 2,778 SNPs for which at least one of the longitudinal tests had a P-value $< 10^{-4}$, we were able to retrieve 2,602 SNP summary statistics from Yale-Penn. Pathway analysis was conducted on the top 84 SNPs that had a P-value < 0.05. Although none of the top twelve SNPs in Table 1.2 had a P-value < 0.05 in the Yale-Penn AA sample, each of the two pathways remained significant: the opioid signaling pathways ($P = 5.67 \times 10^{-4}$, adjusted $P = 3.77 \times 10^{-3}$) and the axonal guidance signaling ($P = 2.89 \times 10^{-4}$, adjusted $P = 2.97 \times 10^{-3}$).

1.4.4 Computation Time

The computational burden of the two retrospective tests, L-BRAT and RGMMAT, mainly comes from the eigendecomposition of the GRM in calculating the retrospective variance of the score functions. However, its impact on run time is minimal because the decomposition needs to be done only once per genome scan. When fitting the null

models, the GLMM-based methods require extra time to obtain the estimates of random effects compared to the GEE-based methods. Once the null model is obtained, the transformed phenotypic residual vector, $\hat{\boldsymbol{\Gamma}}_0^{1/2}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\Gamma}}_0^{-1/2}(\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0)$, in L-BRAT and the phenotypic residual vector, $\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_0$, in RGMMAT, need to be calculated just once per genome scan. Thus, the computational cost of the variance in the retrospective tests is much less than that in the prospective tests. We reported some example run times for analysis of simulated and real data. For a simulated dataset of phenotypes at five time points on 2,000 individuals, the GEE-based methods took 0.9 s while the GLMM-based methods took 37 s to fit the null model. Overall, L-BRAT took 2.4 s and GEE took 27.7 s to analyze 1,000 SNPs using a single processor on an Intel Xeon 2.6 GHz CPU machine. In the analysis of the VACS cocaine use data, L-BRAT and GEE took 1 s while RGMMAT and GMMAT took 2.5 min to fit the null model. Overall, L-BRAT, RGMMAT, GEE, and GMMAT took 0.8 hr, 0.7 hr, 24.8 hr, and 26.2 hr, respectively, to analyze a total of 10,215,072 genome-wide SNPs on Intel Xeon 2.6 GHz CPU computing clusters with 22 nodes. These results demonstrate that L-BRAT and RGMMAT are computationally feasible for large-scale whole-genome association studies.

1.5 Discussion

Longitudinal data can be used in GWAS to improve power for identification of genetic variants and environmental factors that influence complex traits over time. In this study, we have developed L-BRAT, a retrospective association testing method for longitudinal binary outcomes. L-BRAT is based on GEE, thus it requires assumptions on

the mean but not the full distribution of the outcome. Correct specification of the covariance of repeated measurements within each individual is not required, instead, a working covariance matrix is assumed. The significance of the L-BRAT association test is assessed retrospectively by considering the conditional distribution of the genotype at the variant of interest, given phenotype and covariate information, under the null hypothesis of no association. Features of L-BRAT include the following: (1) it is computationally feasible for genetic studies with millions of variants, (2) it allows both static and time-varying covariates to be included in the analysis, (3) it allows different individuals to have measurements at different time points, and (4) it has correct type I error in the presence of ascertainment and trait model misspecification. For comparison, we also propose a retrospective, logistic mixed model-based association test, RGMMAT, which requires specification of the full distribution of the outcome. Random effects are used to model dependence among observations for an individual. Like L-BRAT, RGMMAT is a retrospective analysis in which genotypes are treated as random conditional on the phenotype and covariates. As a result, RGMMAT is also more robust to misspecification of the model for the phenotype distribution than GMMAT test.

Through simulation, we demonstrated that the type I error of L-BRAT was well calibrated under different trait models and ascertainment schemes, whereas the type I error of the prospective GEE method was inflated relative to nominal levels. In the GLMM-based methods, GMMAT, a prospective analysis, was overly conservative, whereas the retrospective version, RGMMAT, was able to maintain correct type I error. We further demonstrated that the two retrospective tests, L-BRAT and RGMMAT, provided higher power to detect association than the prospective GEE and GMMAT tests

under all the trait models and ascertainment schemes considered in the simulations. The choice of the working correlation matrix in L-BRAT resulted in little loss of power. We applied L-BRAT and RGMMAT to longitudinal association analysis of cocaine use in the VACS data, where we identified six novel genes that are associated with cocaine use. Moreover, our pathway analysis identified two significant pathways associated with longitudinal cocaine use: the opioid signaling pathway and the axonal guidance signaling pathway. We were able to replicate both pathways in a cocaine dependence case-control GWAS from the Yale-Penn study. Lastly, we illustrated that the top SNPs identified by our methods are more likely to be the amygdala eQTLs in the GTEx data. The amygdala plays an important role in the processing of memory, decision-making, and emotional responses, and contributes to drug craving that leads to addiction and relapse [41,42]. These findings verify that L-BRAT is able to detect important loci in a genome scan and to provide novel insights into the disease mechanism in relevant tissues. Both simulation studies and the real data analysis suggest that, in general, L-BRAT is a more robust and at the same time, computationally more efficient test than RGMMAT.

Although both L-BRAT and RGMMAT are proposed for population samples, they can be easily extended to related samples in family data for whom the pedigree structure is known. Use the similar strategy of CERAMIC, which extends the CARAT to related samples, it would allow us to incorporate the partially missing data to enhance power. To extend L-BRAT to family design, we could include the kinship matrix into the correlation structure and also modify the genotypic model to incorporate the possibility that genotype being related to covariates. Also, we should consider a more robust variance estimator that incorporating kinship matrix into estimation.

The L-BRAT and RGMMAT methods are designed for single-variant association analysis of longitudinally measured binary outcomes. However, single-variant association tests in general have limited power to detect association for low-frequency or rare variants in sequencing studies. We have previously developed longitudinal burden test and sequence kernel association test, LBT and LSKAT, to analyze rare variants with longitudinal quantitative phenotypes [19]. Both tests are based on a prospective approach. To extend L-BRAT and RGMMAT to rare variant analysis with longitudinal binary data, we could consider either a linear statistic or a quadratic statistic that combines the retrospective score test at each variant in a gene region. In addition, the genetic effect in L-BRAT and RGMMAT is assumed to be constant. We could consider an extension to allow for time-varying genetic effect so that the fluctuation of genetic contributions to the trait value over time is well calibrated.

Chapter 2

Variant-set Retrospective Association Tests for Longitudinal Traits

2.1 Abstract

Longitudinal repeated measures have been increasingly used in genome-wide association studies. The repeated measures provide an opportunity to study the temporal development of traits and also increase the statistical power in association tests. Most of the existing variants-set association tests are based on a population model in which ascertainment sampling is ignored. Prospective inference with longitudinal traits and rare variants can have inflated type I error when the trait model is misspecified. Here, we propose LSRAT (Longitudinal variant-Set Retrospective Association Tests) and RSMMAT (Retrospective variant-Set Mixed Model Association Tests), two groups of retrospective variant-set tests that are constructed based on the genotype model given the phenotype and covariates. RSMMAT can be viewed as a retrospective version of the recently proposed variant-set mixed model association tests (SMMAT) and the LSRAT tests are derived under the generalized estimation equation framework. These two retrospective tests are robust against trait model misspecification and are computationally more efficient than existing prospective approaches. Simulation studies showed that our proposed tests are robust to the trait model misspecification and gain power compared to

SMMAT. We illustrated our method in the Veterans Aging Cohort Study to evaluate the association of repeated measures of alcohol use disorder with rare variants.

2.2 Introduction

Longitudinally measured phenotypes where each individual has multiple follow-ups over the time are more available in genome-wide association studies (GWAS). With the emergence of electronic health record (EHR) in large long-term studies such as UK Biobank and Million Veteran Program, longitudinal measures are vastly being introduced to genomic studies and GWAS. Compared to analysis based on single-time-point measures or traits values averaged over time, longitudinal measures make full use of phenotype information which renders more powerful genetic association tests. Moreover, analyzing longitudinal measures enables the incorporation and adjusting for time-varying covariates into the model. It also delivers the opportunity to study temporal development of complex traits. Because of these merits, an increasing number of studies have developed association tests using longitudinally measured phenotypes in GWAS [3,4,14,43–45]. However, most studies have only focused on single variant tests.

Regardless of the extensive discovery of the genetic common variants associated with complex traits, the identified genetic variants explain only a fraction of total heritability which is often termed “missing heritability”. Although many possible explanations have been proposed, one of the most widely accepted is that the additional heritability can be found by studying rare variants which have the minor allele frequency (MAF) of less than 5% [46]. However, single variant tests for rare variants are underpowered due to the

extremely low MAF. In this regard, many various variant-set association tests have been developed to aggregate multiple rare variants within a region, for example, within a gene or a biological pathway, to increase the detection power. Among such tests, the most two popular approaches are burden tests and the Sequence Kernel Association Test (SKAT)[47]. Burden tests consider a weighted sum of multiple genetic variants into a single score and are powerful when the effects of those variants in a group are homogeneity in direction and magnitude [48]. However, when the genomic region of interest contains signal of both directions (e.g., both risk and protective effects), burden tests may lose power. By contrast, SKAT evaluates the variance of the genetic effects of a group of genetic variants by adopting a statistic of quadratic form. It is more robust to regions where variants' effects are in opposite directions [49] and allows different directions and magnitudes of signals. There are several omnibus tests that unify both the burden and SKAT tests and borrow strength from both approaches, for example, the SKAT-O[47], MiST[50], aSPU[51], SMMAT-E[52], and ACAT-O [53]. Recently, an aggregated Cauchy association test (ACAT) has been proposed which efficiently combines Cauchy transformed p-values. The set-based ACAT (ACAT-V) which combines variant-level p-values has been shown to have strong power when the genetic association signal is sparse; and the omnibus ACAT (ACAT-O) which combines multiple set-based tests provides another strategy to combine SKAT and burden statistics. However, all these variant-set association tests were developed for single-time-point measures and are not directly applicable to longitudinal repeated measures. Thus, there is a pressing need to develop powerful and efficient variant-set tests for rare variants in longitudinal GWA studies.

There are some attempts to fill this need. For example, Wang et al. extended the burden and the SKAT statistics for longitudinal continuous phenotypes by introducing L-Burden and L-SKAT[44]. The two tests were developed under the linear mixed model framework that takes into account interpersonal correlation. Similarly, He et al. extended the GEE-based SKAT test for longitudinal continuous traits[54]. However, these tests were derived from prospective models, and therefore rely on the correct specification of phenotype model to maintain correct type I errors. Yet for longitudinal dichotomous traits, especially for rare diseases and conditions, efficient sampling is widely used in which subjects are usually sampled based on their baseline measures. Prospective analysis in which a population-based model is used overlooks the ascertainment bias and may cause compromised statistical inference[10]. To overcome this limitation, several retrospective association methods have been proposed to analyze ascertained case-control studies [10,12]. Contrast to prospective analysis, retrospective approaches in which genotypes are treated as random conditional on phenotype and covariates are robust to phenotype model misspecification and ascertainment bias. Recently, we extended the retrospective tests to the study of longitudinal binary traits by proposing L-BRAT (GEE-based) and RGMMAT (GLMM-based). However, both L-BRAT and RGMMAT are single-variant test designed for common genetic variant and thus there still lacks retrospective approaches for longitudinal variant-set association test.

Here, we propose LSRAT (Longitudinal variant-Set Retrospective Association Tests) and RSMMAT (Retrospective variant-Set Mixed Model Association Tests), two groups of longitudinal retrospective variant-set tests that are constructed based on the genotype model given the phenotype and covariates. RSMMAT can be viewed as a retrospective

version of the recently proposed variant-set mixed model association tests (SMMAT) and the LSRAT tests are derived under the generalized estimation equation (GEE) framework. These tests have several advantages: (1) they are robust against trait model misspecification; (2) they are able to adjust both static and time-varying covariates; (3) they allow for related subjects and account for population structure; and (4) they are computationally more efficient than existing prospective approaches. Simulation studies showed that our proposed tests are robust to the trait model misspecification and gain power compared to SMMAT and GEE tests. We illustrated our method in the Veterans Aging Cohort Study to evaluate the association of repeated measures of alcohol use with rare variants.

2.3 Methods

We consider the problem of association testing between a set of variants in a genetic region and a longitudinal trait. Suppose genotype, phenotype, and covariate data on a sample of n individuals are available. The genotype data consist of genotypes at the m variants to be tested. The phenotype data consist of repeated measurements of a continuous or binary trait. The covariates are allowed to have both static variables such as sex and dynamic variables such as age. We let n_i denote the number of phenotype measures on individual i and $N = \sum_{i=1}^{n} n_i$ denote the total number of observations. For the ith individual, let Y_{ij} be the trait value, continuous or binary, and $\boldsymbol{X}_{ij}$ be the p-dimensional covariate vector including an intercept, measured at time t_{ij}. We define $\boldsymbol{Y} = (Y_{1,1}, \dots, Y_{1,n_1}, \dots, Y_{n,1}, \dots, Y_{n,n_n})^T$, the trait vector of length N, and $\boldsymbol{X} =$

$(X_{1,1}, \ldots, X_{1,n_1}, \ldots, X_{n,1}, \ldots, X_{n,n_n})^T$, the $N \times p$ covariate matrix. Let $\boldsymbol{G}$ denote the $n \times m$ matrix of genotypes, where $G_{ik} = 0$, 1, or 2 is the number of minor alleles of individual i at the kth variant. Here the genotype matrix $\boldsymbol{G}$ is indexed by individual rather than by measurement. In order to match the dimensions of $\boldsymbol{Y}$ and $\boldsymbol{X}$, we consider a vertically expanded genotype matrix $\boldsymbol{BG}$ that maps the genotype data $\boldsymbol{G}$ from the individual level to the measurement level, where $\boldsymbol{B}$ is defined as an $N \times n$ design matrix representing the measurement clustering structure, and its (l, i)th entry B_{li} is an indicator that the lth entry of $\boldsymbol{Y}$ belongs to the measurements on individual i.

In what follows, we introduce three groups of association testing methods for longitudinal traits: (1) GEE-based prospective association tests; (2) LSRAT (Longitudinal variant-Set Retrospective Association Tests); and (3) RSMMAT (Retrospective variant-Set Mixed Model Association Tests). Like the SMMAT tests [52], the GEE-based association tests are prospective analyses in which $\boldsymbol{Y}$ is treated as random conditional on $\boldsymbol{G}$ and $\boldsymbol{X}$, whereas LSRAT and RSMMAT are retrospective analyses in which $\boldsymbol{G}$ is treated as random conditional on $\boldsymbol{Y}$ and $\boldsymbol{X}$.

2.3.1 Generalized Estimating Equations (GEEs)

In the GEE-based analysis, we model the mean of the phenotype distribution, given the genotypes and covariates, as follows

$$E\left(Y_{ij} \mid \boldsymbol{G}, \boldsymbol{X}\right) = \mu_{ij}, \;\; g\left(\mu_{ij}\right) = \boldsymbol{X}_{ij}^T \boldsymbol{\alpha} + \boldsymbol{G}_i^T \boldsymbol{\beta}, \;\; i = 1, \ldots, n; \; j = 1, \ldots, n_i, \qquad (1)$$

where $\boldsymbol{\alpha}$ is an unknown p-dimentional vector of covariate effects, $\boldsymbol{\beta}$ is an unknown m-dimensional vector of genotype effects, and $g(\,)$ is a link function, for example,

$g(\mu_{ij}) = \mu_{ij}$ for continuous phenotypes, and $g(\mu_{ij}) = \text{logit}(\mu_{ij})$ for binary phenotypes. The covariance matrix of $\boldsymbol{Y}$, denoted by $\boldsymbol{\Omega}$, is specified as

$$\text{Var}(\boldsymbol{Y} \mid \boldsymbol{G}, \boldsymbol{X}) = \boldsymbol{\Omega} = \phi \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{1/2},$$

where $\boldsymbol{\Gamma} = \text{diag}\{v(\mu_{1,1}), \ldots, v(\mu_{1,n_1}), \ldots, v(\mu_{n,1}), \ldots, v(\mu_{n,n_n})\}$ is an N-dimentional diagonal matrix, $v(\,)$ is the variance function, with $v(\mu_{ij}) = 1$ for continuous traits and $v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$ for binary traits, $\boldsymbol{\Sigma}$ is an $N \times N$ working correlation matrix which may depend on some parameter δ, and $\phi > 0$ is a dispersion parameter, with $\phi = \sigma^2$ for continuous phenotypes and $\phi = 1$ for binary phenotypes. The working correlation matrix $\boldsymbol{\Sigma}$ is allowed to be misspecified. In Model (1), the genotype effects $\boldsymbol{\beta}$ are assumed to follow a distribution with mean $\boldsymbol{W}\mathbf{1}_m\beta_0$ and covariance $\tau \boldsymbol{W}\boldsymbol{W}$, where $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_m)$ is a fixed, prespecified m-dimentional diagonal weight matrix, $\mathbf{1}_m$ is an m-vector of 1's, and τ is the variance component of genotype effects. The weights $w_1, \ldots, w_m$ specify how the genotype effects depend on particular features of the variants. Various weighting schemes are available, such as uniform weighting, weighting based on some function of the minor allele frequency (MAF) of the variants [49,55], and function or annotation-based weighting.

The GEEs for the unknown parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are constructed as

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}(\boldsymbol{\alpha}) \\ \boldsymbol{U}(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{\Delta} \boldsymbol{\Omega}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \\ (\boldsymbol{B}\boldsymbol{G})^T \boldsymbol{\Delta} \boldsymbol{\Omega}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \end{bmatrix},$$

where $\boldsymbol{\Delta} = \text{diag}\{u(\mu_{1,1}), \ldots, u(\mu_{1,n_1}), \ldots, u(\mu_{n,1}), \ldots, u(\mu_{n,n_n})\}$ is an N-dimentional diagonal matrix and $u(\,)$ is the first derivative of $g^{-1}(\,)$. To detect association between

the trait and the genetic region of interest, we test $H_0: \boldsymbol{\beta} = \mathbf{0}$ vs. $H_1: \boldsymbol{\beta} \neq \mathbf{0}$. The score function for the genotype effects $\boldsymbol{\beta}$ under H_0 can be written as

$$\boldsymbol{U}_0(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{\beta})_{|\hat{\boldsymbol{\alpha}}_0, \mathbf{0}, \hat{\delta}_0} = (\boldsymbol{BG})^T \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0).$$

Here $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Delta}}_0$ and $\widehat{\boldsymbol{\Omega}}_0$ are $\boldsymbol{\mu}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$ evaluated at $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta) = \left(\hat{\boldsymbol{\alpha}}_0, \mathbf{0}, \hat{\delta}_0\right)$, where $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\delta}_0$ are the null estimates of $\boldsymbol{\alpha}$ and δ under the constraint $\boldsymbol{\beta} = \mathbf{0}$, which can be computed iteratively between a Fisher scoring algorithm for $\boldsymbol{\alpha}$ and the method of moments for δ until convergence.

Hypothesis Testing: GEE-B, GEE-S, GEE-O, GEE-E, GEE-C, and GEE-A

In the GEE model of Eq. (1), our primary interest is to test the genotype effects $H_0: \boldsymbol{\beta} = \mathbf{0}$, which is equivalent to test the null hypothesis $H_0: \beta_0 = 0$ and $\tau = 0$. If we assume $\tau = 0$ and test the null hypothesis $H_0: \beta_0 = 0$, the GEE-based burden test GEE-B has the form

$$T_{GEE-B} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0 \boldsymbol{BGW} \mathbf{1}_m \mathbf{1}_m^T \boldsymbol{W} \boldsymbol{G}^T \boldsymbol{B}^T \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0). \quad (2)$$

Under the null hypothesis, the statistic T_{GEE-B} is asymptotically distributed as $\varphi_B \chi_1^2$, where χ_1^2 is a chi-squared distribution with 1 df, the scalar $\varphi_B = \mathbf{1}_m^T \boldsymbol{W} \boldsymbol{G}^T \boldsymbol{B}^T \boldsymbol{Q} \boldsymbol{BGW} \mathbf{1}_m$ and $\boldsymbol{Q} = \boldsymbol{\Lambda} - \boldsymbol{\Lambda X} (\boldsymbol{X}^T \boldsymbol{\Lambda X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\mathrm{Cov}}(\boldsymbol{Y}) \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0$ and the sample covariance of $\boldsymbol{Y}$, $\widehat{\mathrm{Cov}}(\boldsymbol{Y})$, is estimated by $(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T$.

If we assume $\beta_0 = 0$ and test $H_0: \tau = 0$, the GEE-based variance component SKAT test GEE-S has the form

$$T_{GEE-S} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0 \boldsymbol{BGWW} \boldsymbol{G}^T \boldsymbol{B}^T \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0). \quad (3)$$

Under the null hypothesis, T_{GEE-S} asymptotically follows $\sum_{k=1}^{m} \varphi_{S_k} \chi_{1,k}^2$, where $\chi_{1,k}^2$ are independent chi-squared distributions with 1 df, and φ_{S_k} are the eigenvalues of the matrix $\boldsymbol{WG^T B^T QBGW}$.

Like SKAT-O, we can combine the SKAT and burden tests by considering

$$T_{GEE-O} = \pi T_{GEE-B} + (1-\pi) T_{GEE-S}.$$

We can see that T_{GEE-O} reduces to the GEE burden test when $\pi = 1$ and to the GEE SKAT test when $\pi = 0$. An optimal π can be chosen from the data by minimizing the p value of T_{GEE-O}, following a similar approach to the SKAT-O method[47].

An alternative joint test, similar to MiST [50] and SMMAT-E [52] that were designed under the mixed effects models, for testing $H_0: \beta_0 = 0$ and $\tau = 0$ can be constructed as two independent tests: (1) a test for $H_0: \beta_0 = 0$ under the constraint $\tau = 0$, and (2) a test for $H_0: \tau = 0$ without any constraint on β_0. The first test is the GEE burden statistic T_{GEE-B}, and the second test T_τ can be constructed as

$$T_\tau = (\boldsymbol{Y} - \widetilde{\boldsymbol{\mu}})^T \widetilde{\boldsymbol{\Omega}}^{-1} \widetilde{\boldsymbol{\Delta}} \boldsymbol{BGWWG^T B^T} \widetilde{\boldsymbol{\Delta}} \widetilde{\boldsymbol{\Omega}}^{-1} (\boldsymbol{Y} - \widetilde{\boldsymbol{\mu}}),$$

where $\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Delta}}$ and $\widetilde{\boldsymbol{\Omega}}$ are $\boldsymbol{\mu}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$ evaluated at $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta) = \left(\widetilde{\boldsymbol{\alpha}}, \boldsymbol{W}\mathbf{1}_m \tilde{\beta}_0, \tilde{\delta}\right)$. Here $\widetilde{\boldsymbol{\alpha}}$, $\tilde{\beta}_0$ and $\tilde{\delta}$ are the estimates of $\boldsymbol{\alpha}$, β_0 and δ under a burden-type of mean model

$$E\left(Y_{ij} \mid \boldsymbol{G}, \boldsymbol{X}\right) = \mu_{ij}, \;\; g\left(\mu_{ij}\right) = \boldsymbol{X}_{ij}^T \boldsymbol{\alpha} + \boldsymbol{G}_i^T \boldsymbol{W}\mathbf{1}_m \beta_0, \;\; i = 1, \ldots, n; \; j = 1, \ldots, n_i. \qquad (4)$$

We can show that

$$T_\tau = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)^T \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0 \boldsymbol{BGW} \Big\{ \mathbf{I}_m - \mathbf{1}_m \big(\mathbf{1}_m^T \boldsymbol{WG}^T \boldsymbol{B}^T \widehat{\boldsymbol{P}} \boldsymbol{BGW} \mathbf{1}_m \big)^{-1} \mathbf{1}_m^T \boldsymbol{WG}^T \boldsymbol{B}^T \widehat{\boldsymbol{P}} \boldsymbol{BGW} \Big\} \Big\{ \mathbf{I}_m - \boldsymbol{WG}^T \boldsymbol{B}^T \widehat{\boldsymbol{P}} \boldsymbol{BGW} \mathbf{1}_m \big(\mathbf{1}_m^T \boldsymbol{WG}^T \boldsymbol{B}^T \widehat{\boldsymbol{P}} \boldsymbol{BGW} \mathbf{1}_m \big)^{-1} \mathbf{1}_m^T \Big\} \boldsymbol{WG}^T \boldsymbol{B}^T \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0),$$

where $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0 - \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0 \boldsymbol{X} \big(\boldsymbol{X}^T \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0 \boldsymbol{X} \big)^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{\Delta}}_0 \widehat{\boldsymbol{\Omega}}_0^{-1} \widehat{\boldsymbol{\Delta}}_0$.

Some assumptions: Since the true value of β_0 small, we assume including the genetic burden score in Eq. (4) doesn't dramatically change the variance function matrix $\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}$. Independence holds when the working correlation matches the true correlation.

Alternatively, one can apply ACAT approach to combine P values of individual tests. The variant-set ACAT test can be written as

$$T_{GEE-C} = \sum_{k=1}^{m} w_k' f(p_k)$$

where p_k is the P-value of score test of the k[th] genetic variants; $w_k' = w_k \sqrt{MAF_k(1 - MAF_k)}$ specifies the weight for the P-value which depends on the minor allele frequency of genetic variant; $f(x) = tan\{(0.5 - x)\pi\}$ performs Cauchy transformation on each of the P-values. And the omnibus ACAT test can be used to combined the above three variant-level tests

$$T_{GEE-A} = \frac{1}{3}[f(p_{GEE-B}) + f(p_{GEE-S}) + f(p_{GEE-C})]$$

Both T_{GEE-C} and T_{GEE-A} P-values can be calculated via the Cauchy-distribution-based approximation.

In contrast to SMMAT [52] and the GEE-based association tests which are prospective analyses based on the phenotype given the genotypes and covariates, LSRAT and RSMMAT are retrospective analyses based on the genotypes given the phenotype and covariates. The advantage of the retrospective approach is that the inference is robust to misspecification of the phenotype model, i.e., the type I error of the association tests is still properly controlled given correct specification of the null conditional mean and variance of the genotype data, but not the phenotype model. Since LSRAT and RSMMAT have very similar forms, for clarity, we first present the retrospective model and the test statistics for LSRAT, and then briefly describe the RSMMAT statistics and emphasize the differences between the two retrospective analyses.

2.3.2 LSRAT Model and test statistics

In LSRAT, we specify a retrospective mean model of the genotype

$$E(\boldsymbol{G}_k \,|\boldsymbol{Y}, \boldsymbol{X}) = 2p_k\mathbf{1}_n + \gamma_k\boldsymbol{\Phi}\boldsymbol{A}, \;\; k = 1, \dots, m, \tag{5}$$

where $\boldsymbol{G}_k$ is the genotype vector at the kth variant, p_k is its MAF which is treated as an unknown nuisance parameter, γ_k is an unknown parameter of interest representing the strength and direction of association between the phenotype and the kth variant, $\mathbf{1}_n$ is an n-vector of 1's, $\boldsymbol{\Phi}$ is an $n \times n$ genetic relationship matrix (GRM) representing the overall genetic similarity between individuals due to population structure, and $\boldsymbol{A}$ is an individual-level transformed phenotypic residual vector, where we let $\boldsymbol{A} = \boldsymbol{B}^T\widehat{\boldsymbol{\Delta}}_0\widehat{\boldsymbol{\Omega}}_0^{-1}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)$, obtained from the null GEE model $g(\mu_{ij}) = \boldsymbol{X}_{ij}^T\boldsymbol{\alpha}$, for $i = 1, \dots, n;\; j = 1, \dots, n_i$. If we let

$\widetilde{\boldsymbol{G}} = \mathrm{vec}(\boldsymbol{G}) = (\boldsymbol{G}_1^T, \dots, \boldsymbol{G}_m^T)^T$ be an nm-dimentional vector denoting the vectorization of the genotype matrix $\boldsymbol{G}$, Model (5) can be equivalently written as

$$E\left(\widetilde{\boldsymbol{G}} \,\middle|\, \boldsymbol{Y}, \boldsymbol{X}\right) = 2\boldsymbol{p} \otimes \mathbf{1}_n + \boldsymbol{\gamma} \otimes \boldsymbol{\Phi}\boldsymbol{A}, \tag{6}$$

where $\boldsymbol{p} = (p_1, \dots, p_m)^T$ is a vector of the MAFs of the m genetic variants, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m)^T$ is an unknown vector of association parameters of interest, and $\otimes$ is Kronecker product.

To form LSRAT, we require the null conditional covariance matrix of $\widetilde{\boldsymbol{G}}$, which can be specified as

$$\mathrm{Var}_0\left(\widetilde{\boldsymbol{G}} \,\middle|\, \boldsymbol{Y}, \boldsymbol{X}\right) = \boldsymbol{\Sigma}_G \otimes \boldsymbol{\Phi}, \tag{7}$$

where $\boldsymbol{\Sigma}_G = \boldsymbol{D}^{1/2}\boldsymbol{R}\boldsymbol{D}^{1/2}$ is an $m \times m$ covariance matrix of the variants. Here $\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_m^2)$ is the marginal variance of the m genetic variants, and $\boldsymbol{R}$ is the correlation matrix that captures the linkage disequilibrium (LD) structure. In Model (6), the genotype-phenotype association parameters $\boldsymbol{\gamma}$ are assumed to follow a distribution with mean $\boldsymbol{V}\mathbf{1}_m\gamma_0$ and covariance $\theta\boldsymbol{V}\boldsymbol{V}^T$, where $\boldsymbol{V}$ is a prespecified $m \times m$ weight matrix and θ is the variance component of the association parameters. Note that the weight matrix $\boldsymbol{V}$ in the retrospective model of Eq. (6) plays a similar role as the weight matrix $\boldsymbol{W}$ in the prospective GEE model of Eq. (1) that is allowed to depend on features of the variants. However, we do not require $\boldsymbol{V}$ to be a diagonal or symmetric matrix. In fact, the connection between $\boldsymbol{\gamma}$ of Eq. (6) and $\boldsymbol{\beta}$ of Eq. (1) is that $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_G\boldsymbol{\beta}$ when $\boldsymbol{\beta}$ tends to zero. Therefore, one choice for the weight matrix $\boldsymbol{V}$ in the retrospective model (6) is that $\boldsymbol{V} = \boldsymbol{\Sigma}_G\boldsymbol{W}$. Then, the quasi-likelihood score function for $\boldsymbol{\gamma}$ is given by (see Supplementary Materials S 2.1)

$$\boldsymbol{U}(\boldsymbol{\gamma}) = \boldsymbol{\Sigma}_G^{-1}\boldsymbol{G}^T\boldsymbol{A} - (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\gamma}.$$

Hypothesis Testing: LSRAT-B, LSRAT-S, LSRAT-O, LSRAT-E, LSRAT-C, and LSRAT-A

To detect association between the trait and a genetic region of interest, we test $H_0: \boldsymbol{\gamma} = \mathbf{0}$ vs. $H_1: \boldsymbol{\gamma} \neq \mathbf{0}$ in the retrospective model for $\boldsymbol{G}$ conditional on $\boldsymbol{Y}$ and $\boldsymbol{X}$ given in Eq. (6), which is equivalent to test the null hypothesis $H_0: \gamma_0 = 0$ and $\theta = 0$. If we assume $\theta = 0$ and test $H_0: \gamma_0 = 0$, the LSRAT burden statistic LSRAT-B can be constructed as

$$T_{LSRAT-B} = \boldsymbol{U}_0^T(\boldsymbol{\gamma})\boldsymbol{V}\mathbf{1}_m\mathbf{1}_m^T\boldsymbol{V}^T\boldsymbol{U}_0(\boldsymbol{\gamma}) = \boldsymbol{A}^T\boldsymbol{G}\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{G}^T\boldsymbol{A}, \quad (8)$$

where $\boldsymbol{U}_0(\boldsymbol{\gamma}) = \boldsymbol{U}(\boldsymbol{\gamma})_{|\boldsymbol{\gamma}=\mathbf{0}} = \widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{G}^T\boldsymbol{A}$. Under the null model, the covariance of $\boldsymbol{G}^T\boldsymbol{A}$ is $\mathrm{Cov}(\boldsymbol{G}^T\boldsymbol{A}) = (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\boldsymbol{\Sigma}_G$ (see Supplementary Materials S2.2 for details). Then, the statistic $T_{LSRAT-B}$ asymptotically follows $\lambda_B\chi_1^2$, where the scalar $\lambda_B = (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m$ and χ_1^2 is a chi-squared distribution with 1 df.

If we assume $\gamma_0 = 0$ and test $H_0: \theta = 0$, the LSRAT variance component SKAT statistic LSRAT-S can be constructed as

$$T_{LSRAT-S} = \boldsymbol{U}_0^T(\boldsymbol{\gamma})\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{U}_0(\boldsymbol{\gamma}) = \boldsymbol{A}^T\boldsymbol{G}\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{G}^T\boldsymbol{A}. \quad (9)$$

Under the null hypothesis, $T_{LSRAT-S}$ asymptotically follows $\sum_{k=1}^m \lambda_{S_k}\chi_{1,k}^2$, where $\chi_{1,k}^2$ are independent chi-squared distributions with 1 df, and λ_{S_k} are the eigenvalues of the matrix $(\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}$.

The SKAT-O type of test for LSRAT can be formulated as a weighted average of the LSRAT-B and LSRAT-S statistics, given by

$$T_{LSRAT-O} = \pi T_{LSRAT-B} + (1-\pi)T_{LSRAT-S}.$$

An optimal π is obtained through a grid search by minimizing the p value of $T_{LSRAT-O}$.

Analogous to the GEE-E and SMMAT-E tests in which two independent tests were constructed, we can modify the quasi-likelihood score statistics to perform a joint test of the null hypothesis $H_0: \gamma_0 = 0$ and $\theta = 0$. Specifically, we first test $H_0: \gamma_0 = 0$ under the constraint $\theta = 0$, which is the LSRAT burden statistic $T_{LSRAT-B}$, and then test $H_0: \theta = 0$ without any constraint on γ_0. The second variance component test can be constructed from the null retrospective burden model

$$E\left(\widetilde{\boldsymbol{G}} \,\middle|\, \boldsymbol{Y}, \boldsymbol{X}\right) = 2\boldsymbol{p} \otimes \mathbf{1}_n + \boldsymbol{V}\mathbf{1}_m\gamma_0 \otimes \boldsymbol{\Phi}\boldsymbol{A}. \tag{10}$$

If we assume the mean of association effects γ_0 is small and including the second term in Eq. (10) does not change the conditional covariance of $\widetilde{\boldsymbol{G}}$, we obtain the estimate of γ_0, denoted by $\tilde{\gamma}_0$, by solving the quasi-likelihood score equation

$$U(\gamma_0) = (\boldsymbol{V}\mathbf{1}_m)^T[\boldsymbol{\Sigma}_G^{-1}\boldsymbol{G}^T\boldsymbol{A} - (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\boldsymbol{\Sigma}_G^{-1}\boldsymbol{V}\mathbf{1}_m\gamma_0] = 0,$$

given by

$$\tilde{\gamma}_0 = \frac{\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{G}^T\boldsymbol{A}}{(\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})(\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m)}.$$

Then, the quasi-likelihood score function for $\boldsymbol{\gamma}$ under Model (9) is

$$\boldsymbol{U}_B(\boldsymbol{\gamma}) = \boldsymbol{U}(\boldsymbol{\gamma})_{|\tilde{\gamma}_0} = \widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{G}^T\boldsymbol{A} - (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m\tilde{\gamma}_0$$

$$= \left[\widehat{\boldsymbol{\Sigma}}_G^{-1} - \widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m\left(\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m\right)^{-1}\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\right]\boldsymbol{G}^T\boldsymbol{A} = \boldsymbol{P}\boldsymbol{G}^T\boldsymbol{A},$$

where $\boldsymbol{P} = \widehat{\boldsymbol{\Sigma}}_G^{-1} - \widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m\left(\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}\boldsymbol{V}\mathbf{1}_m\right)^{-1}\mathbf{1}_m^T\boldsymbol{V}^T\widehat{\boldsymbol{\Sigma}}_G^{-1}$. Finally, the variance component statistic T_θ can be written as

$$T_\theta = \boldsymbol{U}_B^T(\boldsymbol{\gamma})\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{U}_B(\boldsymbol{\gamma}) = \boldsymbol{A}^T\boldsymbol{G}\boldsymbol{P}\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{P}\boldsymbol{G}^T\boldsymbol{A}.$$

Under the null hypothesis $H_0: \theta = 0$, T_θ is asymptotically distributed as $\sum_{k=1}^{m} \lambda_{\theta_k} \chi_{1,k}^2$, where $\chi_{1,k}^2$ are independent chi-squared distributions with 1 df, and λ_{θ_k} are the eigenvalues of the matrix $(\boldsymbol{A}^T \boldsymbol{\Phi} \boldsymbol{A}) \boldsymbol{V}^T \boldsymbol{P} \boldsymbol{V}$. By the central limit theorem, both $\boldsymbol{V}^T \boldsymbol{P} \boldsymbol{G}^T \boldsymbol{A}$ and $\mathbf{1}_m^T \boldsymbol{V}^T \widehat{\boldsymbol{\Sigma}}_G^{-1} \boldsymbol{G}^T \boldsymbol{A}$ asymptotically follow normal distributions, and their covariance is

$$\mathrm{Cov}\left(\boldsymbol{V}^T \boldsymbol{P} \boldsymbol{G}^T \boldsymbol{A}, \mathbf{1}_m^T \boldsymbol{V}^T \widehat{\boldsymbol{\Sigma}}_G^{-1} \boldsymbol{G}^T \boldsymbol{A}\right) = (\boldsymbol{A}^T \boldsymbol{\Phi} \boldsymbol{A}) \boldsymbol{V}^T \boldsymbol{P} \boldsymbol{V} \mathbf{1}_m = \mathbf{0}.$$

Therefore, $T_{LSRAT-B}$ and T_θ are asymptotically independent. We use Fisher's method to combine the p values from the two tests.

Note that all the above four LSRAT statistics involve $\boldsymbol{G}^T \boldsymbol{A}$, the product of the column vectors in the genotype matrix $\boldsymbol{G}$ and the phenotypic residual vector $\boldsymbol{A}$, where $\boldsymbol{A}$ is obtained from the null prospective GEE model of the phenotype. As we can alternatively generate the phenotypic residuals based on the GLMM model of the phenotype, we propose in the next section a group of retrospective association testing methods using the phenotypic residuals obtained from the GLMM.

The retrospective version of variant-set ACAT test and omnibus ACAT test are proposed as follows:

$$T_{LSRAT-C} = \sum_{k=1}^{m} w_k' f(p_k^r)$$

and

$$T_{LSRAT-A} = \frac{1}{3}\left[f(p_{LSRAT-B}) + f(p_{LSRAT-S}) + f(p_{LSRAT-C})\right]$$

where p_k^r is the L-BRAT P-value on the k^{th} genetic variant (e.g., retrospective single-variant test).

2.3.3 RSMMAT Model and Test Statistics

The SMMAT tests [52] were formulated from the GLMM

$$g(\mu_{ij}) = \boldsymbol{X}_{ij}^T\boldsymbol{\alpha} + \boldsymbol{G}_i^T\boldsymbol{\beta} + b_{ij},\ \ i = 1, \dots, n;\ j = 1, \dots, n_i,$$

where $\mu_{ij} = E\left(Y_{ij} \mid \boldsymbol{G_i}, \boldsymbol{X}_{ij}, b_{ij}\right)$ is the mean of a response at time t_{ij} for individual i, given his/her genotypes, covariates, and random effect, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the same as defined in Model (1), the vector of random effects $\boldsymbol{b} = (b_{1,1}, \dots, b_{1,n_1}, \dots, b_{n,1}, \dots, b_{n,n_n})^T$ is assumed that $\boldsymbol{b} \sim MVN(\boldsymbol{0}, \sum_{l=1}^{L} \nu_l\, \boldsymbol{\Phi}_l)$ with L variance component parameters ν_l, and correlation matrices $\boldsymbol{\Phi}_l$. Here we allow for multiple random effects to capture the correlation among repeated measures from longitudinal studies.

Fitting the null GLMM model $g(\mu_{ij}) = \boldsymbol{X}_{ij}^T\boldsymbol{\alpha} + b_{ij}$, for $i = 1, \dots, n;\ j = 1, \dots, n_i$, we generate an n-dimensional vector of phenotypic residuals $\boldsymbol{C}$ at the individual level, defined by $\boldsymbol{C} = \boldsymbol{B}^T(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_0)/\hat{\phi}$, where $\widehat{\boldsymbol{\mu}}_0 = g^{-1}(\boldsymbol{X}\widehat{\boldsymbol{\alpha}}_0 + \widehat{\boldsymbol{b}})$ is a vector of fitted values, and $\hat{\phi}$ is an estimate of the dispersion parameter ϕ.

Different from LSRAT, the RSMMAT model specifies that

$$E\left(\widetilde{\boldsymbol{G}} \,\middle|\, \boldsymbol{Y}, \boldsymbol{X}\right) = 2\boldsymbol{p} \otimes \boldsymbol{1}_n + \boldsymbol{\gamma} \otimes \boldsymbol{\Phi C}, \tag{11}$$

and the null conditional covariance matrix of $\widetilde{\boldsymbol{G}}$ is the same as that in Eq. (7), then the quasi-likelihood score function for $\boldsymbol{\gamma}$ is written as $\boldsymbol{U}(\boldsymbol{\gamma}) = \boldsymbol{\Sigma}_G^{-1}\boldsymbol{G}^T\boldsymbol{C} - (\boldsymbol{C}^T\boldsymbol{\Phi C})\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\gamma}$. Finally, we construct the four RSMMAT statistics RSMMAT-B, RSMMAT-S, RSMMAT-O, and RSMMAT-E by replacing the phenotypic residual vector $\boldsymbol{A}$ in the corresponding LSRAT tests with the GLMM-based phenotypic residual vector $\boldsymbol{C}$; RSMMAT-C and RSMMAT-A by replacing the GEE score tests with GLMM ones. Their distributions can be similarly obtained.

2.3.4 Connection between Retrospective and Prospective Tests

To assume $\boldsymbol{\beta}$ has mean $\boldsymbol{W}\mathbf{1}_m\beta_0$ and variance $\tau\boldsymbol{W}\boldsymbol{W}$, it would be equivalent to assume $\boldsymbol{\gamma}$ has mean $\boldsymbol{\Sigma}_G\boldsymbol{W}\mathbf{1}_m\beta_0$ and variance $\tau\boldsymbol{\Sigma}_G\boldsymbol{W}\boldsymbol{W}\boldsymbol{\Sigma}_G$. If we define $\boldsymbol{V} = \boldsymbol{\Sigma}_G\boldsymbol{W}$, $\gamma_0 = \beta_0$ and $\theta = \tau$, then we will have $T_B^r = T_B$ and $T_S^r = T_S$. Hence, we show the connection between retrospective burden and SKAT statistics derived from retrospective mean model and the original burden and SKAT test statistics. Moreover, to perform retrospective burden and SKAT testing, it is equivalent to evaluating the distribution of burden and SKAT score statistics retrospectively.

In this section, we introduce a retrospective model, and show the connection between prospective and retrospective statistics.

2.4 Simulation Studies

2.4.1 Simulation of Type I Error

We performed extensive simulations to examine the type I error of LSRAT-B, S, C, O, E, A and RSMMAT-B, S, C, O, E, A, and compare their empirical power with that of GEE and the GLMM tests. For all the simulations, we generated 10,000 chromosomes over a 1Mb regions using a coalescent model and mimicking the LD structure, the recombination rate, and the population history of the European population. We generated sequence data with 100 genetic variants selected from 4kb region in each set and 1000,000 independent sets for 7,500 individuals with seven repeated measures.

For continuous traits, in each simulation replicate, we simulated the phenotype y_{ij} for subject i's jth observation under the null hypothesis without genetic effects from

$$y_{ij} = 1.0 + 0.5X_{1,i} + 0.5X_{2,ij} + \beta_{time}(j-1) + a_i + r_{ij} + e_{ij}$$

$$i = 1, \ldots n; j = 1, \ldots 7$$

where $X_{1,i}$ is a continuous time-varying covariate that is generated from independently from a standard normal distribution; $X_{2,ij}$ is a binary time-invariant covariate with the probability of taking value 1 of 0.5; $\beta_{time} = 2.0$ models time effects; a_i and r_{ij} are the individual-level time-independent and time-dependent random effects, respectively. We assume $a_i \sim N(0, \sigma_a^2)$ and $r_i = (r_{i1}, \ldots r_{i7}) \sim MVN(0, \sigma_r^2 \boldsymbol{R})$, where $\boldsymbol{R}$ is a 7×7 correlation matrix specified by the AR(1) structure with a correlation coefficient ψ. The parameters for the variance components are set $\sigma_a^2 = \sigma_r^2 = \sigma_e^2 = 0.64$ and correlation coefficient $\psi = 0.7$.

For the binary traits, in each simulation replicate, we simulated the phenotype y_{ij} for subject i's jth observation under the null hypothesis without genetic effects from

$$\text{logit}\left(P(y_{ij} = 1)\right) = -2.5 + 0.5X_{1,i} + 0.5\, X_{2,ij} + \beta_{time}(j-1) + a_i + r_{ij}$$

where $\beta_{time} = 0.2$ represents time effects on the probability of developing the disease. All other parameters are the same as those in the continuous traits. We consider a baseline ascertainment scheme for the longitudinal dichotomous trait where 3750 case and 3750 control subjects were sampled according to their outcome values at baseline.

2.4.2 Simulation of Empirical Power

To assess the power performance of comparing set-based tests, we randomly selected causal variants within each of the genetic regions to simulate phenotypes under the alternatives. Specifically, we generated continuous longitudinal phenotype by

$$y_{ij} = 1.0 + 0.5X_{1,i} + 0.5X_{2,ij} + \beta_{time}(j-1) + \sum_{q=1}^{s} G_{iq}\beta_q + a_i + r_{ij} + e_{ij}$$

and dichotomous longitudinal measures by

$$\text{logit}\left(P(y_{ij}=1)\right) = -2.5 + 0.5X_{1,i} + 0.5\,X_{2,ij} + \beta_{time}(j-1) + \sum_{q=1}^{s} G_{iq}\beta_q + a_i + r_{ij}$$

where $G_1, \ldots G_s$ are the genotypes of randomly selected causal variants β_qs are the effect sizes for the causal variants, and the other symbols are the same as defined in the simulation for type I error.

To investigate the impact of causal proportion, effect direction and sample sizes on the power of different tests, we vary the above factors in our power simulation studies. The proportion of causal variants was set to be 5%, 20% and 50% which covers cases of sparse and dense signals. The causal effect directions of positive/negative directions were set to be 50/50%, 80%/20% and all positive to represent different mixture proportions of protective and deleterious causal variants. The effect size ($|\beta_q|$s) was set to be $c|\log_{10} MAF_q|$, such that variants with a smaller MAF have a large effect size, where the c depends on the causal proportion. We examined three sample size designs: 2,500, 5,000 and 7,500. Similar to the type I error simulation, we performed baseline ascertainment for each of the sample size designs for the longitudinal dichotomous traits. We repeated this procedure 1,000 times to obtain P-values for the power estimation for each test.

2.4.3 Simulation Results

Table 2.1 shows the empirical type I error rates of LSRAT- B, S, C, O, E, A and RGMMAT- B, S, C, O, E, A and their prospective GEE tests counterparts at significance levels of 0.01, 0.001, and 0.0001 in the variant set analysis of continuous traits. All

twenty-four tests have well-controlled type I error rates at these significance levels except GEE-C, which has slightly conservative type I error. Table 2.2 shows the empirical type I error rate of the above twenty-four tests in the variant set analysis of baseline ascertained dichotomous traits at significance level of 0.01, 0.001, and 0.0001. All twenty-four tests have well-controlled type I error rates at theses significance levels for dichotomous traits.

Table 2.1. Type I Error Estimates for Each Tests Aimed that Testing the Association Between Randomly Selected Genetics Regions with a Continuous Longitudinal Traits. The sample size is 7,500 subjects with seven repeated measure and type I error rate under the basis of 10^6 replicates.

		alpha = 0.01		alpha = 0.001		alpha = 0.0001	
		GEE	LSRAT	GEE	LSRAT	GEE	LSRAT
GEE-based	Burden(B)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	1.1E-04	1.0E-04
	SKAT(S)	9.0E-03	1.0E-02	9.0E-04	1.0E-03	9.0E-05	1.0E-04
	ACAT-V(C)	8.0E-03	1.0E-02	6.0E-04	1.0E-03	6.0E-05	1.1E-04
	SKAT-O(O)	1.0E-02	1.1E-02	1.1E-03	1.2E-03	1.1E-04	1.2E-04
	SMMAT-E(E)	9.0E-03	1.0E-02	9.0E-04	1.0E-03	1.0E-04	1.1E-04
	ACAT-O(A)	9.0E-03	1.1E-02	8.0E-04	1.0E-03	8.0E-05	1.1E-04
		GLMM	RSMMAT	GLMM	RSMMAT	GLMM	RSMMAT
GLMM-based	Burden(B)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	1.0E-04	1.0E-04
	SKAT(S)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	9.0E-05	9.0E-05
	ACAT-V(C)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	9.0E-05	1.0E-04
	SKAT-O(O)	1.1E-02	1.1E-02	1.2E-03	1.2E-03	1.1E-04	1.1E-04
	SMMAT-E(E)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	1.1E-04	1.1E-04
	ACAT-O(A)	1.1E-02	1.1E-02	1.1E-03	1.1E-03	9.0E-05	9.0E-05

Table 2.2. Type I Error Estimates for Each Tests Aimed that Testing the Association Between Randomly Selected Genetics Regions with a Baseline Ascertained Dichotomous Longitudinal Traits. The sample size is 7,500 subjects with seven repeated measure and type I error rate under the basis of 10^6 replicates.

		alpha = 0.01		alpha = 0.001		alpha = 0.0001	
		GEE	LSRAT	GEE	LSRAT	GEE	LSRAT
GEE-based	**Burden(B)**	1.0E-02	1.0E-02	1.0E-03	1.0E-03	9.0E-05	1.0E-04
	SKAT(S)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	1.0E-04	1.1E-04
	ACAT-V(C)	1.0E-02	1.0E-02	1.0E-03	1.1E-03	8.0E-05	1.2E-04
	SKAT-O(O)	1.1E-02	1.1E-02	1.1E-03	1.2E-03	1.3E-04	1.4E-04
	SMMAT-E(E)	1.0E-02	1.0E-02	1.0E-03	1.0E-03	1.0E-04	1.1E-04
	ACAT-O(A)	1.0E-02	1.1E-02	1.0E-03	1.1E-03	9.0E-05	1.2E-04
		GLMM	**RSMMAT**	**GLMM**	**RSMMAT**	**GLMM**	**RSMMAT**
GLMM-based	**Burden(B)**	1.0E-02	1.0E-02	9.0E-04	1.0E-03	7.0E-05	8.0E-05
	SKAT(S)	1.0E-02	1.0E-02	9.0E-04	1.0E-03	8.0E-05	9.0E-05
	ACAT-V(C)	9.0E-03	1.0E-02	9.0E-04	1.0E-03	1.1E-04	1.2E-04
	SKAT-O(O)	1.0E-02	1.0E-02	9.0E-04	1.0E-03	9.0E-05	1.0E-04
	SMMAT-E(E)	1.0E-02	1.1E-02	1.1E-03	1.1E-03	1.3E-04	1.4E-04
	ACAT-O(A)	9.0E-03	1.0E-02	9.0E-04	9.0E-04	9.0E-05	1.0E-04

We compare the power between LSRAT and GEE, RSMMAT and SMMAT, respectively under a variety of simulation conditions for both continuous and dichotomous traits. Figure 2.1 presents the empirical power of LSRAT- B, S, C, O, E, A for testing causal variant sets evaluated at the significance level of 2.5×10^{-6} for longitudinal continuous traits. Each LSRAT test was compared with GEE- B, S, C, O, E, A, respectively. Figure 2.2 presents the empirical power of LSRAT tests and GEE tests for longitudinal dichotomous traits evaluated at the same significance level. LSRAT tests have improved power compared to their corresponding GEE tests for longitudinal

continuous traits, especially for longitudinal dichotomous traits. Among them, LSRAT-V, A, O has the most substantial power gain as compared with GEE-V, A, O.

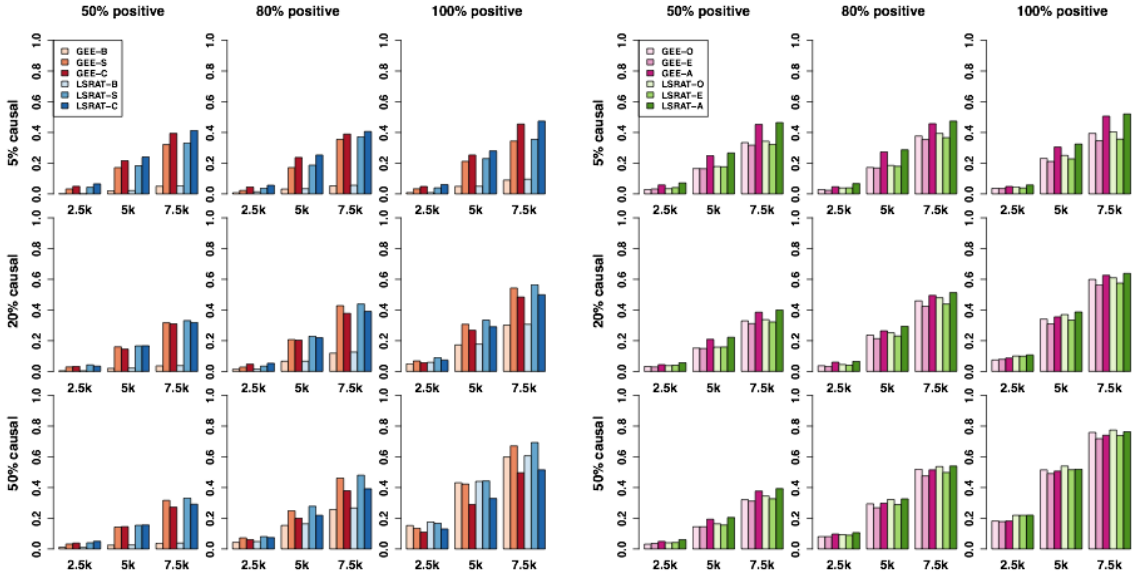


Figure 2.1. Power plots of variant-set tests(left panel) GEE-B, S, C and LSRAT-B, S, C; omnibus tests(right panel) GEE-O, E, A and LSRAT- O, E, A for continuous longitudinal traits. Each bar represents the empirical power estimated as the proportion of p values less than $\alpha = 5 \times 10^{-6}$ of sample size $n = 2{,}500$, 5,000, 7,500. The proportion of causal variants is set to be 5%, 20%, and 50% which were shown by three rows of each panel. The coefficients for the causal variants are 50% positive, 80% positive, and 100% positive which corresponds to the three columns of each panel. The effect size($|\beta_i|$ s) of the causal variants are set to be $|\beta_i| = c|\log_{10} MAF_i|$, where c was set to 0.04 for 50% causal, 0.06 for 20% causal and 0.12 for 5% causal.

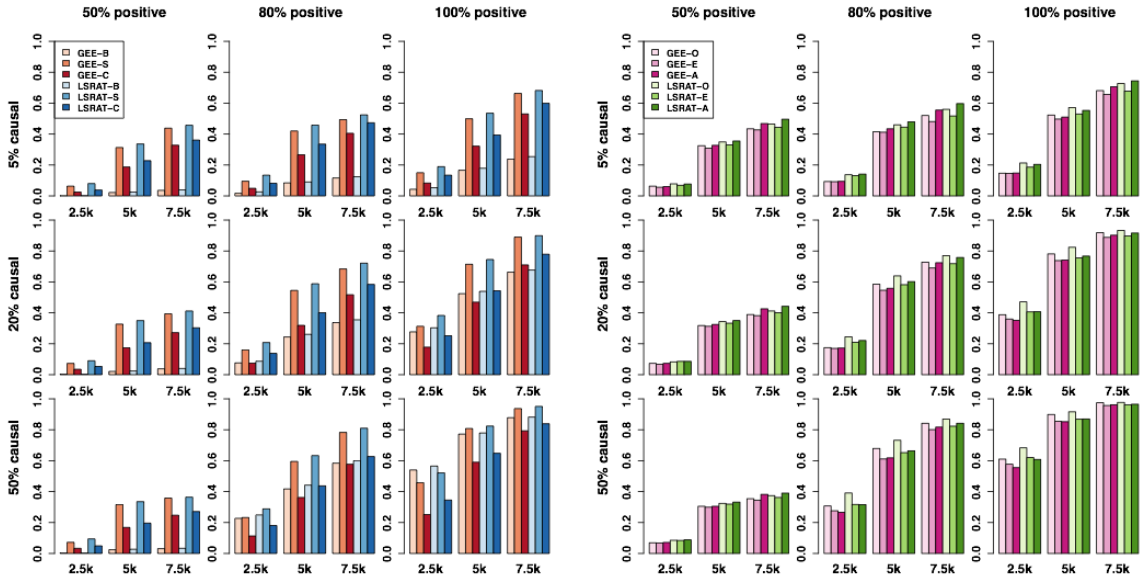


Figure 2.2. Power plots of variant-set tests(left panel) GEE-B, S, C and LSRAT-B, S, C; omnibus tests(right panel) GEE-O, E, A and LSRAT- O, E, A for dichotomous longitudinal traits. Each bar represents the empirical power estimated as the proportion of p values less than $\alpha = 5 \times 10^{-6}$ of sample size $n = 2{,}500$, 5,000, 7,500. The proportion of causal variants is set to be 5%, 20%, and 50% which were shown by three rows of each panel. The coefficients for the causal variants are 50% positive, 80% positive, and 100% positive which corresponds to the three columns of each panel. The effect size($|\beta_i|$ s) of the causal variants are set to be $|\beta_i| = c|\log_{10} MAF_i|$, where c was set to 0.08 for 50% causal, 0.12 for 20% causal and 0.24 for 5% causal.

Figure 2.3 and figure 2.4 present the empirical power of RSMMAT- B, S, C, O, E, A for testing causal variant sets evaluated at the significance level of 2.5×10^{-6} for longitudinal continuous and dichotomous traits. We compared their power with their prospective counterparts SMMAT B, S, C, O, E, A, respectively. For longitudinal continuous traits, because the prospective model is correctly specified, retrospective GLMM-based tests have similar power as prospective ones. For longitudinal dichotomous traits in which subjects were ascertained based on their baseline observations, the prospective modeling is misspecified. In this situation, RSMMAT tests have substantially increased power as compared with GLMM tests. The most notable increase of power is observed comparing the aggregated Cauchy association variant-set

test (RSMMAT-C and GLMM-C), where GLMM-C has substantially lesser power than RSMMAT-C. This is because the aggregated Cauchy association variant-set test is based on the P-values of each single variant test. And as was shown in the previous study[56], the GLMM-based single variant test is underpowered in ascertained phenotypes, which compromised the power of the aggregated test.

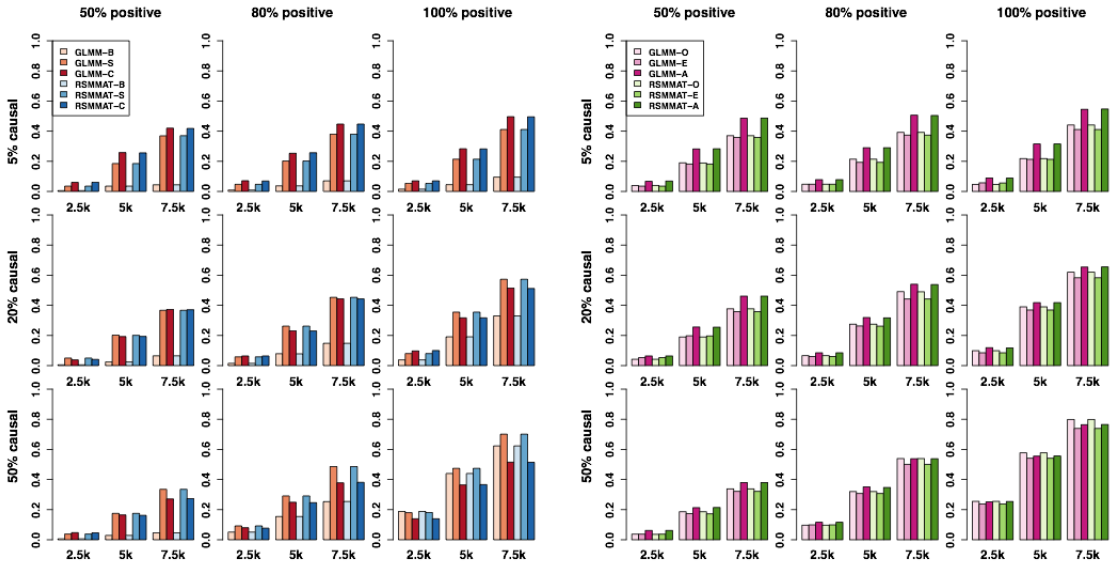


Figure 2.3. Power plots of variant-set tests (left panel) GLMM-B, S, C and RSMMAT-B, S, C; omnibus tests (right panel) GLMM-O, E, A and RSMMAT- O, E, A for continuous longitudinal traits. Each bar represents the empirical power estimated as the proportion of p values less than $\alpha = 5 \times 10^{-6}$ of sample size $n = 2{,}500$, 5,000, 7,500. The proportion of causal variants is set to be 5%, 20%, and 50% which were shown by three rows of each panel. The coefficients for the causal variants are 50% positive, 80% positive, and 100% positive which corresponds to the three columns of each panel. The effect size($|\beta_i|$ s) of the causal variants are set to be $|\beta_i| = c|\log_{10} MAF_i$where c was set to 0.04 for 50% causal, 0.06 for 20% causal and 0.12 for 5% causal.

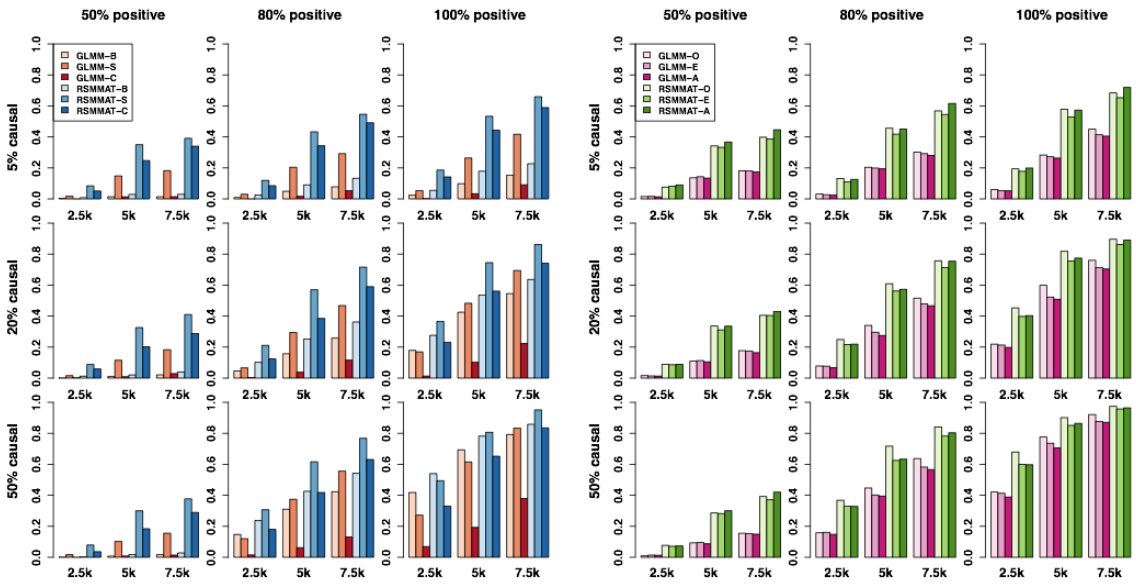


Figure 2.4. Power plots of variant-set tests (left panel) GLMM-B, S, C and RSMMAT-B, S, C; omnibus tests(right panel) GLMM -O, E, A and RSMMAT- O, E, A for dichotomous longitudinal traits. Each bar represents the empirical power estimated as the proportion of p values less than $\alpha = 5 \times 10^{-6}$ of sample size $n = 2{,}500$, 5,000, 7,500. The proportion of causal variants is set to be 5%, 20%, and 50% which were shown by three rows of each panel. The coefficients for the causal variants are 50% positive, 80% positive, and 100% positive which corresponds to the three columns of each panel. The effect size($|\beta_i|$ s) of the causal variants are set to be $|\beta_i| = c|\log_{10} MAF_i|$, where c was set to 0.08 for 50% causal, 0.12 for 20% causal and 0.24 for 5% causal.

The power increases with the sample size and decreases when the proportion of causal signals of the same direction drops from 100% to 50% for all tests, but with the most considerable decrease observed for burden types of tests. Among the three types of variant-set tests, the ACAT-V tests are more powerful in the 5% causal scenario where the causal signal is sparse but less powerful when the signal is dense (20% and 50% causal). Burden tests are more powerful when the causal proportion is high (50%) and the signals are in the same direction (100% positive) for a small sample size (2.5k), but the advantage of burden tests diminishes as the sample size increases. For omnibus tests, all three types of the omnibus tests in general show similar power and robustness to the

direction and proportion of causal variants. SKAT-O has an advantage when the there is a large set of causal variants (50%) while ACAT-O gains advantage when the proportion of causal variants is small (5%). SKAT-E has slightly less powerful than SKAT-O but it is computationally more efficient.

Comparing the performance of GEE-based tests and GLMM-based tests, the simulation results suggest that GLMM-based tests are slightly more powerful than GEE-based tests for longitudinal continuous traits. This is due to the fact that the GLMM model fitted has correctly specified random components whereas GEE model used AR1 correlation structure and sandwich estimator. On the other hand, GEE-based tests are much more powerful than GLMM-based tests for longitudinal dichotomous traits with efficient sampling. This suggests that GEE-based tests are more robust to ascertainment sampling than GLMM-based tests. Additionally, the retrospective analysis demonstrates more powerful gain than prospective tests in longitudinal dichotomous traits.

2.5 Application: VACS Alcohol Use GWAS Data

2.5.1 Association Analysis

To illustrate the use of our proposed tests, we analyzed a GWAS data set of alcohol use disorder from VACS [57]. VACS is a longitudinal observational cohort study of both HIV-positive and uninfected veterans. It has the aim of understanding the role of psychiatric conditions including alcohol and other substance abuse in the clinical consequences of HIV infection. Our use of the VACS data was approved by both the Yale Human Research Protection Program and the Institutional Review Board of the

Veterans Affairs Connecticut Healthcare System. Our data source was alcohol use disorders identification test-consumption (AUDIT-C) questionnaire. This data is longitudinal as the questionnaire was collected over time for six clinic visits, on a total of 2,470 patients. The AUDIT-C score is a reliable and valid measure to assess the risk of harmful alcohol use, which has been used in previous VACS studies. The score ranges 0 to 12, where 0 reflects no alcohol use and evaluates three measures of alcohol consumption. These measures include the frequency of usual consumption, the quantity of usual consumption, and the frequency of binge drinking. The missing rate at each visit ranges from 3.0% to 58.3%. The average AUDIC-C score for HIV positive subjects is 3.90 and for HIV negative subjects is 4.19.

All samples were genotyped on the Illumina OmniExpress BeadChip and were imputed using 1000 Genome Phase 3 data as a reference panel using IMPUTE2[23]. We performed quality-control (QC) on the subjects and the genetic variants. A detailed description of the subject level QC process can be found in our previous study[56]. SNPs that satisfied all of the following QC conditions were included in this analysis: (1) call rate > 95%, (2) Hardy-Weinberg χ^2 statistic P-value > 10^{-6}. We annotated variants to genes using ANNOVAR [58].The resulting data set has 2,210 individuals who have both genotype and phenotype information and a total of 32,233 genes. Gender, age at baseline, HIV status, and top five principal components (PCs) were included as static covariates and the time of visit was included as a dynamic covariant.

We performed genome-wide association gene-based testing for the longitudinal AUDIT-C score on 32,233 genes in a total of 2,210 subjects. We considered GEE, LSRAT, GLMM and RSMMAT tests with adjustment for sex, baseline age, HIV status,

visit time and top five PCs. For each of the four types of tests (GEE, GLMM, LSRAT, RSMMAT), we consider three variant-set tests (B, S, C) and three omnibus tests (O, E, A). In total, we evaluated twenty-four tests. The genomic control inflation factors for the twelve tests ranges from 0.92 to 1.06.

Table 2.3 summarizes genes identified by GEE model based prospective and retrospective variant-set and omnibus tests (GEE and LSRAT) with P-values less than 5×10^{-5} in at least one longitudinal test. Table 2.4 summarizes genes identified by GLMM model based prospective and retrospective variant-set and omnibus tests (GLMM and RSMMAT) with P-values less than 10^{-4} in at least one longitudinal test. Gene UBE2L3 was identified as a top gene from both GEE model based tests and GLMM model based tests. This gene, encoding ubiquitin-conjugating enzyme E2, was recently reported in another GWA study to interact with alcohol consumption and was also significantly associated with lipid levels [59]. In a previous study, it was identified as an ethanol-responsive gene in the prefrontal cortex in mice[60]. This suggests that UBE2L3 could be implicated in alcoholism. Gene EFCAB10, located at 7q22.3, is another protein coding gene that was also identified among top genes in both models. An SNP located within this gene was formerly found in association with bipolar disorder [61], a commonly co-occurred disorder with alcoholism [62]. Gene NPIPB3 is a protein coding gene located at 16p12.2. This gene was reported to be associated with extremely impulsively violence and aggressive behavior in males. As was suggested in previous studies, the frequently observed comorbidity of alcohol use disorder and impulsive aggression implicate a shared genetic basis underlies these two disorders [63–65]. In another study, an intron variant (rs12923444) located in the promoter region of NPIPB3

was found significantly associated with depression in a Genome-wide meta-analysis[66]. Our tests found gens to be significantly associated with alcohol use disorder that have been previously shown to be associated with alcoholism or major psychiatric disorders that comorbid with alcoholism.

Table 2.3. Top genes with P-value < 5×10^{-5} in at least one of the GEE and LSRAT in the VACS sample. * denotes protein coding gene, bold denotes the minimum P-value for the given gene. The smallest P-value among all tests at the given genes are in bold.

			Prospective						Retrospective					
Gene	SNPs	Chr	GEE-B	GEE-S	GEE-C	GEE-O	GEE-E	GEE-A	LSRAT-B	LSRAT-S	LSRAT-C	LSRAT-O	LSRAT-E	LSRAT-A
MIR4454	10	4	4.49E-04	1.44E-05	4.65E-05	3.25E-05	6.95E-05	3.15E-05	1.28E-03	8.73E-05	3.32E-04	1.61E-04	3.48E-04	1.98E-04
UBE2L3*	324	22	3.52E-05	2.63E-03	1.50E-02	6.11E-05	2.43E-04	7.52E-05	1.97E-05	4.08E-04	1.02E-02	5.46E-05	1.73E-04	5.89E-05
C9orf40*	15	9	8.34E-02	7.90E-02	1.03E-03	1.30E-01	3.68E-02	4.57E-03	8.00E-02	6.81E-02	2.21E-05	1.28E-01	2.36E-02	7.70E-05
LINC01455	136	5	7.31E-02	5.55E-03	6.10E-03	1.23E-02	1.89E-02	9.14E-03	7.18E-02	2.90E-03	4.29E-05	6.22E-03	3.09E-03	1.45E-04
ZNF33A*	202	10	8.58E-02	9.30E-04	5.04E-05	2.16E-03	2.33E-03	2.54E-04	7.85E-02	4.71E-04	8.40E-04	1.09E-03	2.57E-03	9.24E-04
ZNF25*	130	10	4.17E-01	6.41E-02	5.45E-05	9.97E-02	1.31E-02	3.07E-04	3.82E-01	4.79E-02	9.66E-04	7.71E-02	1.54E-02	3.17E-03
TMEM198B	14	12	5.58E-04	3.83E-02	5.51E-02	9.62E-04	2.62E-03	1.31E-03	5.93E-05	1.06E-02	2.27E-02	1.72E-04	5.08E-04	1.87E-04
NPIPB4*	8	16	6.85E-05	8.43E-03	1.38E-02	1.41E-04	5.64E-04	1.51E-04	8.65E-05	7.45E-03	2.14E-02	1.99E-04	8.78E-04	2.70E-04
EGF*	483	4	5.86E-01	3.72E-04	6.98E-05	9.84E-04	7.40E-04	2.76E-04	6.13E-01	2.44E-04	7.82E-04	6.36E-04	5.21E-04	5.68E-04
PROX1-AS1	233	1	6.63E-01	2.68E-02	1.02E-02	5.45E-02	4.07E-02	2.69E-02	6.66E-01	2.12E-02	8.64E-05	4.19E-02	2.43E-02	2.96E-04
EFCAB10*	8	7	1.29E-01	5.20E-04	1.83E-03	1.40E-03	1.18E-03	1.21E-03	9.56E-02	8.86E-05	1.30E-04	2.47E-04	1.28E-04	1.62E-04

Table 2.4. Top genes with P-value < 10^{-4} in at least one of the GLMM and RSMMAT in the VACS sample. * denotes protein coding gene, bold denotes the minimum P-value for the given gene. The smallest P-value among all tests at the given genes are in bold.

			Prospective						Retrospective					
Gene	SNPs	Chr	GEE-B	GEE-S	GEE-C	GEE-O	GEE-E	GEE-A	LSRAT-B	LSRAT-S	LSRAT-C	LSRAT-O	LSRAT-E	LSRAT-A
UBE2L3*	324	22	3.59E-05	1.32E-03	1.40E-02	9.44E-05	3.31E-04	1.13E-04	5.27E-05	1.09E-03	1.31E-02	1.41E-04	4.46E-04	1.58E-04
LINC01455	136	5	9.13E-02	4.66E-03	6.55E-05	9.76E-03	7.27E-03	2.48E-04	9.80E-02	4.58E-03	3.84E-05	9.66E-03	6.04E-03	1.40E-04
PROX1-AS1	233	1	6.86E-01	1.48E-02	8.44E-05	2.83E-02	1.53E-02	3.20E-04	6.81E-01	1.56E-02	4.55E-05	3.10E-02	1.83E-02	1.66E-04
LINC00467	219	1	9.58E-05	3.11E-03	1.29E-02	2.57E-04	7.81E-04	2.95E-04	6.53E-05	3.03E-03	1.21E-02	1.78E-04	5.53E-04	2.02E-04
C9orf40*	15	9	1.04E-01	1.00E-01	1.33E-04	1.66E-01	4.15E-02	4.98E-04	1.00E-01	1.01E-01	7.22E-05	1.64E-01	4.06E-02	2.63E-04
EFCAB10*	8	7	1.81E-01	1.16E-04	5.59E-04	3.14E-04	2.34E-04	2.97E-04	1.69E-01	9.19E-05	3.99E-04	2.54E-04	1.66E-04	2.24E-04
LOC101927661	2	2	3.68E-02	2.98E-03	1.61E-04	5.43E-03	2.83E-04	5.65E-04	3.36E-02	2.55E-03	1.00E-04	4.60E-03	1.52E-04	3.42E-04

2.5.2 Pathway and eQTL Enrichment Analysis

We performed pathway analysis on the SNPs contained in the top genes for which at least one of the longitudinal tests had a P-value < 5×10^{-5} using METACORE. Of the top ten significant pathways we found, four were particularly of interests. The first one is HTR2A (alias 5-HT2A) signaling, which belongs to neurophysiological process in the nervous system. HTR2A receptor agonist are emerging as a popular therapeutic treatment for alcohol dependence and other neuropsychiatric conditions, and it was studied to normalize dysregulated GABAergic signaling [67]. The second one is canonical WNT signaling pathway, which plays a vital role in neural cell proliferation during neural development. This pathway has been suggested by many studies to be associated with major psychiatric disorders, including bipolar disorder and schizophrenia [68,69]. The third and fourth pathways are adenosine A1 and adenosine A3 receptor signaling. Adenosine plays a crucial role in regulating neural activity in the central nervous system and it modulates many neurotransmitters. It has been found to be central to many pathophysiological processes including drug dependence and alcohol abuse [70,71]. Overall, these four significant pathways all have been shown previously to be associated with major psychiatric disorders and alcoholism, verifying our model is capable of identifying biologically relevant loci.

Next, we performed an enrichment analysis to see if the top genes from our analysis are likely to regulate brain gene expression. Previous studies have shown that genes regulating brain tissue regulation are useful in understating the basis of psychiatric disorders [72]. The cis-eQTLs of 13 human brain regions reported from the GTEx project were considered in the analysis. We performed Fisher's exact test to examine the

enrichment of brain region eQTLs (FDR < 0.05) in the SNPs contained in the top genes from GEE and LSRAT test. Among the brain regions, three brain regions showed significant enrichment: *anterior cingulate cortex* (odds ratio = 9.40, P-val < 2.20×10^{-16}), *cerebellar hemisphere* (odds ratio = 5.53 , P-val = < 2.2×10^{-16}), and *cerebellum*(odds ratio = 3.97 , P-val < 2.20×10^{-16}These three brain regions were also significantly enriched in top genes identified from GLMM and RSMMAT tests: *anterior cingulate cortex* (odds ratio = 4.48, P-val < 2.20×10^{-16}), *cerebellar hemisphere* (odds ratio = 2.26, P-val = 1.01×10^{-10}) and *cerebellum* (odds ratio =1.76 , P-val =1.02×10^{-6}).These results show that the top genes identified from our tests are likely to regulate gene expression in those two brain regions, which will be discussed more extensively in the discussion section.

2.6 Discussion

We have developed and implemented LSRAT and RSMMAT, two families of retrospective variant-set tests for longitudinally measured continuous and binary traits in large scale genome wide association studies. In particular, LSRAT is a family of GEE model based association tests which include three variant-set level tests: the burden test (LSRAT-B), SKAT (LSRAT-S), and ACAT (LSRAT-C), as well as three omnibus tests that combines burden and SKAT with different strategies: LSRAT-O, LSRAT-E, and LSRAT-A. For comparison we also proposed RSMMAT, the mixed model counterparts of LSRAT, which introduced retrospective analysis to the existing prospective variant-set tests (SMMAT). Both LSRAT and RSMMAT are retrospective tests that are constructed

based on the genotype model given the phenotype and covariates. LSRAT models the within subject dependence with working covariance matrix whereas RSMMAT captures it with random effects. These two families of retrospective tests for longitudinal traits have several advantages: (1) they are robust against trait model misspecification; (2) they are able to adjust both static and time-varying covariates; (3) they allow for related subjects and account for population structure; and (4) they are computationally more efficient than existing prospective approaches. They provide important tools for the study of the genetic mechanism of longitudinal phenotypes especially for the psychiatric disorders where the temporal course and developmental pattern of the traits are of valuable information and has been less studied.

LSRAT also has limitations. As mentioned previously, SMMAT p values are computed based on asymptotic distributions, which may be violated in small samples, especially for binary traits. Additionally, the p value computed for SMMAT-E relies on the assumption that the working correlation specified is the true correlation structure. However, benefitting from the robustness of GEE estimators to correlation structure model misspecification, the simulation results showed that LSRAT-E p-value maintained a correct type I error rate even when the correlation structure is not correctly specified.

We applied LSRAT and RSMMAT to longitudinal association analysis of alcohol use in the VACS data. Pathway analysis of the top genes identified four significant pathways associated with longitudinal alcohol use: the HTR2A signaling pathway, the canonical WNT signaling pathway, the adenosine A1 and adenosine A2 signaling pathways. Enrichment analysis of brain region eQTLs demonstrated that top genes comprised of SNPs are enriched with eQTLs from two brain regions: *anterior cingulate cortex* and

cerebellum. The *anterior cingulate cortex* which mediates willed control of actions, was previously found to contribute to drug addiction[73]. Recently, it was studied that the thickness of *anterior cingulate cortex* is associated with alcohol use patterns [74]. As the cerebellar dysfunction and degeneration are commonly observed in alcoholics, the function of which has long been considered to be associated with alcoholism [75]. There are accumulating evidence that connect cerebellum to genetic risk for developing alcohol use disorder. In a recent study, it was demonstrated the strong influence of cerebellar in the susceptibility to alcohol abuse and the cerebellar has been highlighted as a target for pharmacological treatment of alcohol use disorder [76].

In summary, LSRAT and RSMMAT provide a retrospective association framework for variant-set tests in large-scale genome-wide association for longitudinal traits. As the electronic health records become more available, these two families of tests will serve as powerful tools to uncover the mechanism of genes that control the developmental trajectories of traits, especially for psychiatric traits where the progression and developmental trajectories convey more valuable and reliable information than single time points measures. We expect a future extension of the proposed methods towards functional modeling of the genetic temporal effects as well as separately testing gene-environment interaction in longitudinal GWA studies to improve the discovery process of the genetic basis for complex traits.

Chapter 3

Gene-graph based imputation method for single-cell RNA sequencing data

3.1 Abstract

Single-cell RNA sequencing technology provides an opportunity to study gene expression at single-cell resolution. However, prevalent dropout events result in high data sparsity and noise that may obscure downstream analyses in single-cell transcriptomic studies. We propose a novel method, G2S3, that imputes dropouts by borrowing information from adjacent genes in a sparse gene graph learned from gene expression profiles across cells. We applied G2S3 and ten existing imputation methods to eight single-cell transcriptomic datasets to compare their performance. Our results demonstrated that G2S3 is superior in recovering true expression levels, identifying cell subtypes, reconstructing cell trajectories, identifying differentially expressed genes, and recovering gene correlation relationships, especially for genes with relatively low expression levels. Moreover, G2S3 is computationally efficient for imputation in large-scale single-cell transcriptomic datasets.

3.2 Introduction

Singe-cell RNA sequencing (scRNA-seq) has emerged as a state-of-the-art technique for transcriptome analysis. Compared to bulk RNA-seq that measures the average gene

expression profile of a mixed cell population, scRNA-seq measures expression profile of individual cells and thus describes cell-to-cell stochasticity in gene expression. Applications of this technology in humans have revealed rare and novel cell types [77–79], cell population composition changes [80], and cell-type specific transcriptomic changes [79,81] that are associated with diseases. These findings have great potential to promote our understanding of cell function, disease pathogenesis, and treatment response for more precise therapeutic development [82,83]. However, analysis of scRNA-seq data can be challenging due to low library size, high noise level, and prevalent dropout events [84]. Particularly, dropouts lead to an excessive number of zeros or close to zero values in the data, especially for genes with low or moderate expression. These inaccurately measured gene expression levels may obscure downstream quantitative analyses such as cell clustering and differential expression analyses [82].

In the past few years, several imputation methods have been developed to recover dropout events in scRNA-seq data. A group of methods, including kNN-smoothing [85], MAGIC [86], scImpute [87], drImpute [88], and VIPER [89], assess between-cell similarity and impute dropouts in each cell using its similar cells. Specifically, kNN-smoothing uses step-wise k-nearest neighbors to aggregate information from the k closest neighboring cells of each cell for imputation. MAGIC constructs an affinity matrix of cells and aggregates gene expression across similar cells via data diffusion to impute gene expression for each cell [86]. scImpute infers dropout events based on the dropout probability estimated from a Gamma-Gaussian mixture model and only imputes these events by borrowing information from similar cells within cell clusters detected by spectral clustering [87]. drImpute identifies similar cells through K-means clustering and

performs imputation by averaging expression levels of cells within the same cluster [88]. While these imputation methods improved the quality of scRNA-seq data to some extent, they were found to eliminate natural cell-to-cell stochasticity which is an important piece of information available in scRNA-seq data compared to bulk RNA-seq data [89]. VIPER overcomes this limit by considering a sparse set of neighboring cells for imputation to preserve variation in gene expression across cells [89]. In general, imputation methods that borrow information across similar cells tend to intensify subject variation in scRNA-seq datasets with multiple subjects, which causes cells from the same subject to be more similar than those from different subjects. To address this issue, SAVER borrows information across similar genes instead of cells to impute gene expression using a penalized regression model [90]. There are other methods that leverage information from both genes and cells. For example, ALRA imputes gene expression using low-rank matrix approximation [91], and scTSSR uses two-side sparse self-representation matrices to capture gene-to-gene and cell-to-cell similarities for imputation [92]. In addition, machine learning-based methods, such as autoImpute [93], DAC [94], deepImpute [95] and SAUCIE [96], use deep neural network to impute dropout events. While computationally more efficient, these methods were found to generate false-positive results in differential expression analyses [97]. Recently, an ensemble approach, EnImpute, was developed to integrate multiple imputation methods using weighted trimmed mean [98].

In this article, we develop G2S3, a sparse and smooth signal of gene graph-based method that imputes dropout events in scRNA-seq data by borrowing information across similar genes. G2S3 learns a sparse graph representation of gene-gene relationships from

the data, in which each node represents a gene and is associated with a vector of expression levels in all cells considered as a signal on the graph. The graph is then optimized under the assumption that signals change smoothly between connected genes. Based on this graph, a transition matrix for a random walk is constructed so that the transition probabilities between genes with similar expression levels across cells are higher. A random walk on this graph imputes the expression level of each gene using the weighted average of expression levels from itself and adjacent genes in the graph. In this way, G2S3, like SAVER, makes use of gene-gene relationships to recover the true expression levels. However, unlike SAVER which uses a penalized regression model for imputation, G2S3 optimizes the gene graph structure using graph signal processing that captures nonlinear correlations among genes and is robust to outliers in the data. The computational complexity of the G2S3 algorithm is a polynomial of the total number of genes in the graph, so it is computationally efficient, especially for large scRNA-seq datasets with hundreds of thousands of cells.

3.3 Material and methods

3.3.1 G2S3 algorithm

To borrow information from similar genes for data imputation, G2S3 first builds a sparse graph representation of gene network under the assumption that expression levels change smoothly between closely connected genes. Let $X = [x_1, x_2, \dots, x_m] \in \mathbb{R}^{n \times m}$ denote the observed transcript counts of m genes in n cells, where the column $x_j \in \mathbb{R}^n$ represents the expression vector of gene j, for $j = 1, \dots, m$. We regard each gene j as a

vertex V_j in a weighted gene graph $G = (V, E)$, in which the edge between genes j and k is associated with a weight W_{jk}.

The gene graph is then determined by the weighted adjacency matrix $W \in \mathbb{R}_+^{m \times m}$. G2S3 searches for a valid adjacency matrix W from the space

$$\mathcal{W} = \{W \in \mathbb{R}_+^{m \times m}: \ W = W^T, \mathrm{diag}(W) = 0\}$$

that is optimal under the assumption of smoothness and sparsity on the graph. To achieve this, we use the objective function adapted from Kalofolias's model [99]:

$$\min_{W \in \mathcal{W}} \ \|W \circ Z\|_{1,1} - 1^T \log(W1) + \frac{1}{2}\|W\|_F^2, \tag{1}$$

where $Z \in \mathbb{R}_+^{m \times m}$ is the pairwise Euclidean distance matrix of genes, defined as $Z_{jk} = \left\|x_j - x_k\right\|^2$, $\|\cdot\|_{1,1}$ is the elementwise L-1 norm, $\circ$ is the Hadamard product, and $\|\cdot\|_F$ is the Frobenius norm. The first term in Eq. (1) is equivalent to $2\,\mathrm{tr}(X^T L X)$ that quantifies how smooth the signals are on the graph, where L is the graph Laplacian and $\mathrm{tr}(.)$ is the trace of a matrix. This term penalizes edges between distant genes, so it prefers to put a sparse set of edges between the nodes with a small distance in Z. The second term in Eq. (1) represents the node degree which requires the degree of each gene to be positive to improve the overall connectivity of the gene graph. The third term in Eq. (1) controls sparsity to penalize the formation of large edges between genes.

The optimization of Eq. (1) can be solved via primal dual techniques [100]. We rewrite Eq. (1) as

$$\min_{w \in \omega} \ 1_{\{w \geq 0\}} + 2w^T z - 1^T \log(d) + \|w\|^2, \qquad \text{where } \omega = \left\{w \in \mathbb{R}_+^{\frac{m(m-1)}{2}}\right\}, \tag{2}$$

where w and z are vector forms of W and Z, respectively, $d = Kw \in \mathbb{R}^m$ and K is the linear operator that satisfies $W1 = Kw$. After obtaining the optimal W, a lazy random walk matrix can be constructed on the graph:

$$M = (D^{-1}W + I)/2, \tag{3}$$

where D is an m-dimensional diagonal matrix with $D_{jj} = \sum_k W_{jk}$, the degree of gene j, and I is the identity matrix.

The imputed count matrix X_{imputed} is then obtained by taking a t-step random walk on the graph which can be written as

$$X^T_{\text{imputed}} = M^t \times X^T. \tag{4}$$

By default, G2S3 takes a one-step random walk ($t = 1$) to avoid over-smoothing. We also implement an option of tuning the hyperparameter t based on an objective function that minimizes the MSE between the imputed and observed data, i.e.

$$t^* = \underset{t}{\text{argmin}} \ \|M^t X^T - X^T\|.$$

Similar to other diffusion-based methods, G2S3 spreads out counts while keeping the sum constant in the random walk step. This results in the average value of non-zero matrix entry decreasing after imputation. To match the observed expression at the gene level, we rescale the values in X_{imputed} so that the mean expression of each gene in the imputed data matches that of the observed data. The pseudo-code for G2S3 is given in Algorithm 1.

Algorithm 1: Pseudo-code of G2S3

Input: X
Result: $X_{imputed}$ = G2S3(X)
Z = distance(X)
$W = \min_{w \in R_{+}^{m(m-1)/2}} 1_{\{w \geq 0\}} + 2w^T z - 1^T \log(d) + \|w\|^2$
D = degree(W)
$M = (D^{-1} \times W + I)/2$
$t^* = \underset{t}{\text{argmin}} \|M^t X^T - X^T\|$
$X^T_{imputed} = M^{t^*} \times X^T$
$X_{rescaled} = \text{rescale}(X_{imputed})$
$X_{imputed} = X_{rescaled}$
End

3.3.2 Real datasets

We evaluated and compared the performance of G2S3 and ten existing imputation methods using datasets from eight scRNA-seq studies. Among them, four datasets were generated using the UMI techniques and four were generated by non-UMI-based techniques.

Reyfman refers to the scRNA-seq dataset of human lung tissue from healthy transplant donors in Reyfman et al. [101]. The raw data include 33,694 genes and 5,437 cells. To generate the reference dataset, we selected cells with a total number of UMIs greater than 10,000 and genes that have nonzero expression in more than 20% of cells. This ended up with 3,918 genes and 2,457 cells.

PBMC refers to human peripheral blood mononuclear cells from a healthy donor stained with TotalSeq-B antibodies generated by the high-throughput droplet-based system [102]. This dataset was downloaded from 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets). The raw data

include 33,538 genes and 7,865 cells. To generate the reference dataset, we selected cells with a total number of UMIs greater than 5,000 and genes that have nonzero expression in more than 20% of cells. This ended up with 2,308 genes and 2,081 cells.

Zeisel refers to the scRNA-seq dataset of mouse cortex and hippocampus in Zeisel et al. [103]. The raw data include 19,972 genes and 3,005 cells. To generate the reference dataset, we selected cells with a total number of UMIs greater than 10,000 and genes that have nonzero expression in more than 40% of cells. This ended up with 3,529 genes and 1,800 cells.

Chu refers to the dataset investigating the separation of cell subpopulations in Chu et al. [104]. It measured gene expression of 1,018 cells including undifferentiated H1 and H9 human ES cell lines and the H1-derived progenitors. The cells were annotated with seven cell subtypes: neuronal progenitor cells (NP), definitive endoderm cells (DE), endothelial cells (EC), trophoblast-like cells (TB), human foreskin fibroblasts (HF), and undifferentiated H1 and H9 human ES cells. We performed preliminary filtering to remove genes expressed in less than 10% of cells, which resulted in 13,829 genes.

Petropoulos refers to the dataset studying cell lineage in human embryo development in Petropoulos et al. [105]. It measured expression profiles of 26,178 genes in 1,529 cells from 88 human embryos. Cells were labeled as E3-E7 representing their embryonic day. We performed preliminary filtering to remove genes expressed in less than 5 cells and cells with less than 200 expressed genes. After the filtering, we ended up with 22,934 genes and 1,529 cells.

Trapnell refers to the dataset studying the transcriptional dynamics of human myoblasts in Trapnell et al. [106]. scRNA-seq data were collected on undifferentiated

primary human myoblasts at time 0 and differentiating myoblasts at 24, 48 and 72 hours. Most of the cells are mature myotubes 72 hours after inducing differentiation. The raw data include 47,192 genes and 372 cells. We performed preliminary filtering to remove genes expressed in less than 10% of cells, which resulted in 13,286 genes.

Paul refers to the dataset from a study on the transcriptional differentiation landscape of myeloid progenitors [107]. This dataset includes 3,451 informative genes and 2,730 cells. We used this dataset to evaluate the performance of imputation methods in restoring gene regulatory relationships between well-known regulators.

Buettner refers to the dataset in Buettner et al. [108]. This dataset includes mouse ES cells labeled by three cell cycle phases – G1, S, and G2/M via flow sorting. The raw data include 9,571 genes and 288 cells. We used this dataset to evaluate the performance of imputation methods in enhancing gene correlations between periodic marker genes of cell cycle phase. We performed preliminary filtering to remove genes expressed in less than 20% of cells, which resulted in 13,355 genes.

3.3.3 Performance evaluation

Expression data recovery.

We first compared the method performance in recovering true expression levels using down-sampled datasets. Down-sampling was performed on three independent UMI-based scRNA-seq datasets (Reyfman, PBMC, and Zeisel) to generate benchmarking observed datasets in a similar framework to previous studies [90,95]. In each dataset, we selected a subset of genes and cells with high expression to be used as the reference dataset and treated them as the true expression. Details on the thresholds chosen to generate the reference datasets are described in the “Real datasets” section. However, unlike previous studies that

simulated down-sampled datasets from models with certain distributional assumptions [90] which may incur modeling bias, we performed random binary masking of UMIs in the reference datasets to mimic the inefficient capturing of transcripts in dropout events. The binary masking process masked out each UMI independently with a given probability. In each reference dataset, we randomly masked out 80% of UMIs to create the down-sampled observed dataset.

All imputation methods were applied to each down-sampled dataset to generate imputed data separately. Because imputation methods such as SAVER and MAGIC output the normalized library size values, we performed library size normalization on all imputed data. We calculated the gene-wise Pearson correlation and cell-wise Spearman correlation between the reference data and the imputed data generated by each imputation method. The correlations were also calculated between the reference data and the observed data without imputation to provide a baseline for comparison. To investigate whether the performance depends on the true expression level, we stratified genes into three categories: widely, mildly, and rarely expressed genes, based on the proportion of cells expressing each gene in the down-sampled observed datasets. Specifically, widely expressed genes are those with non-zero expression in more than 80% of cells, rarely expressed genes are those with non-zero expression in less than 30% of cells, and mildly expressed genes are those that lie in between. The gene-wise and cell-wise correlations in each stratum were used to demonstrate the impact of expression level on the performance of imputation methods.

Restoration of cell subtype separation.

We applied all imputation methods to the Chu dataset to evaluate their performance in separating different cell types. A good imputation method is expected to stabilize within cell-subtype variation (intra-subtype distance) while maintaining between cell-subtype variation (inter-subtype distance). Principal component analysis was conducted on the raw and imputed data for dimension reduction. We calculated the inter-subtype distance as the Euclidian distance between cells from different cell types, and the intra-subtype distance as the distance between cells of the same cell type, using the top K PCs of the data, for $K = 1, \dots, 50$. The ratio of the average inter-subtype distance to the average intra-subtype distance was used to quantify the performance. The higher this ratio is, the better performance the method has. We also calculated silhouette coefficient, a composite index reflecting both the compactness and separation of different cell types, using the top PCs and the true cell subtype labels. The silhouette coefficient ranges from -1 to 1 with a higher value indicating a better matching with the cell subtypes and a value close to zero indicating random clustering [109]. To demonstrate the comparison using cell clustering results, we visualized the raw and imputed data with UMAP plots using the top three PCs and colored cells by the cell subtype labels. The normalized mutual information (MI) and adjusted rand index (RI) were used to measure the consistency between cell clustering results and true cell subtype labels. To demonstrate cell subtype separation based on cell subtype marker genes, we further displayed DE and H1/H9 cells by plotting the log-transformed counts using their marker genes [104]: *GATA*6, a marker gene of DE cells, and *NANOG*, a marker gene of H1/H9 cells.

Improvement in cell trajectory inference.

We assessed the performance of imputation methods in restoring cell trajectory using human preimplantation embryos from different embryonic days in the Petropoulos dataset. We considered the actual embryonic days to represent the true cell differentiation stage or age. Monocle 2 was used to infer pseudo-time from the normalized raw and imputed data [110]. To measure the consistency between the actual embryonic days and the reconstructed pseudo-time, we calculated the pseudotemporal ordering score (POS) and Kendall rank correlation coefficient (Cor). Cell trajectories were visualized by embedding cells into two-dimensional space using reversed graph embedding, a recently developed machine learning method to reconstruct complex single-cell trajectories [110].

Improvement in differential expression analysis.

To assess the performance in identifying differentially expressed genes, we compared gene expression between two cell subtypes: NP and H1 cells, using both imputed scRNA-seq and bulk RNA-seq data from the Chu dataset. We also compared gene expression profiles of undifferentiated myoblasts to mature myotubes collected 72 hours after inducing differentiation from the Trapnell dataset. The raw and imputed data were normalized and log-transformed before evaluation. We used t-test in the bulk RNA-seq data to identify differentially expressed genes and selected the top 200 genes as ground truth. We then performed differential analysis in the scRNA-seq data using the same test. All the differential expression analysis in the scRNA-seq data was performed using the Seurat R package (version 3.0) with a default threshold to keep genes with at least 1.5-fold change. The predictive power of differentially expressed genes identified in the raw and imputed scRNA-seq data on the ground truth was measured by the area under an ROC curve.

Gene correlation relationship restoration.

We finally evaluated the method performance by investigating the enhancement in gene regulatory relationships using the Paul dataset and the recovery of gene-gene correlations between periodic marker genes in the Buettner dataset. In the Paul dataset, we reconstructed GRN among a set of regulators with known inhibitory and activatory relationships in blood development [94] with the raw and imputed datasets by different methods, using two GRN inference algorithms, GENIE3 and PPCOR. The prediction accuracy of each method was evaluated by comparing the inferred GRN to the ground-truth network using AUPRC. The AUPRC ratio was calculated by dividing AUPRC by that of a random predictor and the process was repeated for 50 times. The estimated pairwise correlations between genes using the raw unimputed and imputed data by each method were compared for performance evaluation. The Beuttner dataset contains 67 periodic marker genes with peak expression in G1/S and G2/M phases established in a previous study [111]. As marker gene expression varies over cell cycle, we expect pairs of periodic genes whose expression peak during the same cell cycle phase to be positively correlated, and pairs of genes whose expression peak at different phases to be negatively correlated. Pairwise correlations were calculated in the raw and imputed data by each method. The proportion of gene pairs with correct direction of correlation was used to compare the method performance.

3.4 Results

3.4.1 Evaluation overview

We evaluated and compared the performance of G2S3 and ten existing imputation methods, SAVER, kNN-smoothing, MAGIC, scImpute, VIPER, ALRA, scTSSR, DCA, SAUCIE and EnImpute, in terms of (1) expression data recovery, (2) cell subtype separation, (3) cell trajectory inference, (4) differential gene identification, and (5) gene regulatory and correlation relationship recovery. We applied these methods to eight scRNA-seq datasets that can be classified into five categories corresponding to the five criteria described above. The first category includes three unique molecular identifier (UMI)-based datasets in which down-sampling was performed to assess the method performance in recovering true expression levels. These datasets are the Reyfman dataset from human lung tissue [101], the peripheral blood mononuclear cell (PBMC) dataset from human peripheral blood [102], and the Zeisel dataset from mouse cortex and hippocampus [103]. The second category was used to evaluate the method performance in separating different cell types. It includes the Chu dataset of human embryonic stem (ES) cell-derived lineage-specific progenitors from seven known cell subtypes [104]. The third category was used to reconstruct cell trajectory. It includes the Petropoulos dataset of cells from human preimplantation embryos collected on different embryonic days [105]. The fourth category was chosen to evaluate the method performance in identifying differentially expressed genes. It includes the Chu dataset which is also included in the second category and the Trapnell dataset of differentiating human myoblasts [106]. The last category includes two datasets to evaluate the method performance in recovering gene regulatory and correlation relationship among known regulators and marker genes. These datasets are the Paul dataset that contains a set of well-known transcriptional regulators of myeloid progenitor populations [107], and the Buettner dataset that contains 67 periodic marker genes whose

expression level varies over cell cycle [108]. Table 1 summarizes the main features of all eight datasets. A more detailed description of these datasets is provided in the "Real datasets" section.

Table 3.1. Detailed information on the eight scRNA-seq datasets used to compare the performance of imputation methods. * URL to access the dataset: https://support.10xgenomics.com/single-cell-gene-expression/datasets

Experiment Category	Dataset	# Cells	Sample Type	Organism	Technique	UMI	Accession
Expression data recovery	Reyfman [23]	5,437	Lung tissue	Homo Sapiens	Drop-seq	Yes	GEO (GSE122960)
	PBMC [24]	7,865	Peripheral blood mononuclear cells	Homo Sapiens	Drop-seq	Yes	10x Genomics*
	Zeisel [25]	3,005	Brain tissue	Mus Musculus	Drop-seq	Yes	Zeisel et al.
Cell subtype separation	Chu [26]	1,018	Embryonic stem cells	Homo Sapiens	Fluidigm C1	No	GEO (GSE75748)
Cell trajectory inference	Petropoulos [27]	1,529	Preimplantation embryos	Homo Sapiens	Smart-seq2	No	Petropoulos et al.
Differential gene identification	Chu [26]	1,018	Embryonic stem cells	Homo Sapiens	Fluidigm C1	No	GEO (GSE75748)
	Trapnell [28]	372	Myoblasts	Homo Sapiens	Fluidigm C1	No	GEO (GSE52529)
Gene correlation relationship recovery	Paul [29]	2,730	Bone marrow myeloid progenitor	Mus Musculus	MARS-seq	Yes	Paul et al.
	Buettner [30]	288	Staged embryonic cells	Mus Musculus	Fluidigm C1	No	ArrayExpress (E-MTAB-2805)

3.4.2 Hyperparameter tuning in G2S3

The G2S3 algorithm used graph signal processing to learn a gene graph and performed a t-step random walk to borrow information from neighboring genes for imputation. The

optimal value of the hyperparameter t was selected by minimizing the mean squared error (MSE) between the imputed and observed data, which was also used in a previous study on diffusion-based imputation method [112]. We performed down-sampling on each dataset from the first category (Reyfman, PBMC and Zeisel) and evaluated the MSE as well as the gene-wise and cell-wise correlations of the G2S3 imputed data with reference data, for $t = 1, \ldots, 10$. Fig S3.1 shows the coefficient of variation (CV) of gene expression before and after down-sampling. In all datasets, although the CV of gene expression increased slightly after down-sampling, the correlation between the CV before and after down-sampling was 0.79 or higher, demonstrating that the down-sampled data well preserved the mean-variance relationship in the reference data. Fig S3.2A shows that the optimal value of t is 1 in all three datasets based on the minimization of MSE. In addition, the one-step random walk in G2S3 achieved the greatest gene-wise and cell-wise correlations with the reference data (Fig S3.2B). This optimal choice of t was consistent with the hyperparameter selected by another diffusion-based imputation method [112].

3.4.3 Expression data recovery in down-sampled datasets

We conducted down-sampling on datasets from the first category (Reyfman, PBMC and Zeisel) to assess the performance of all eleven imputation methods in recovering true expression levels. Fig S3.1 shows the coefficient of variation (CV) of gene expression before and after down-sampling. In all datasets, although the CV of gene expression increased slightly after down-sampling, the correlation between the CV before and after down-sampling was 0.79 or higher, demonstrating that the down-sampled data well preserved the mean-variance relationship in the reference data. Fig 3.1 shows the gene-

wise Pearson correlation and cell-wise Spearman correlation between the imputed and reference data from each dataset. The correlation between the observed data without imputation and reference data was set as a benchmark. In all datasets, G2S3 consistently achieved the highest correlation with the reference data at both gene and cell levels, and SAVER and scTSSR had slightly worse performance. EnImpute had comparable performance to G2S3 based on the cell-wise correlation but performed worse than G2S3, SAVER and scTSSR based on the gene-wise correlation. VIPER performed well in the Reyfman and PBMC datasets but not in the Zeisel dataset based on the gene-wise correlation, although the cell-wise correlations were much lower than G2S3, SAVER, scTSSR and EnImpute in all datasets. The other methods, kNN-smoothing, MAGIC, scImpute, ALRA and DCA, did not have comparable performance, especially based on the gene-wise correlation. SAUCIE did not have comparable performance to the other methods in all datasets (Fig S3.2). Since genes with higher expression tend to have a lower dropout rate, they are usually easier to impute and have less imputation need than those with lower expression [84]. To demonstrate the impact of expression level on the method performance, we stratified genes into three subsets based on the proportion of cells expressing them in the down-sampled data: widely expressed (>80%, n = 155, 111, 110, respectively), mildly expressed (30%-80%, n = 615, 357, 1,902, respectively), and rarely expressed (<30%, n = 3,148, 1,830, 1,617, respectively). Fig S3.3 shows the gene-wise and cell-wise correlations in each gene stratum. We can see that G2S3 improved both gene-wise and cell-wise correlations compared to the observed data for widely and mildly expressed genes. Moreover, G2S3 achieved the most superior recovery accuracy than the other methods for both widely and mildly expressed genes, although SAVER, scTSSR and EnImpute had

comparable accuracy for widely expressed genes, suggesting the advantage of borrowing information from similar genes over from similar cells. For rarely expressed genes, all imputation methods did not improve the correlations compared to the observed data using both gene-wise and cell-wise correlation, suggesting that there is insufficient information for these genes to be successfully imputed. Overall, G2S3 provided the most accurate recovery of true expression levels.

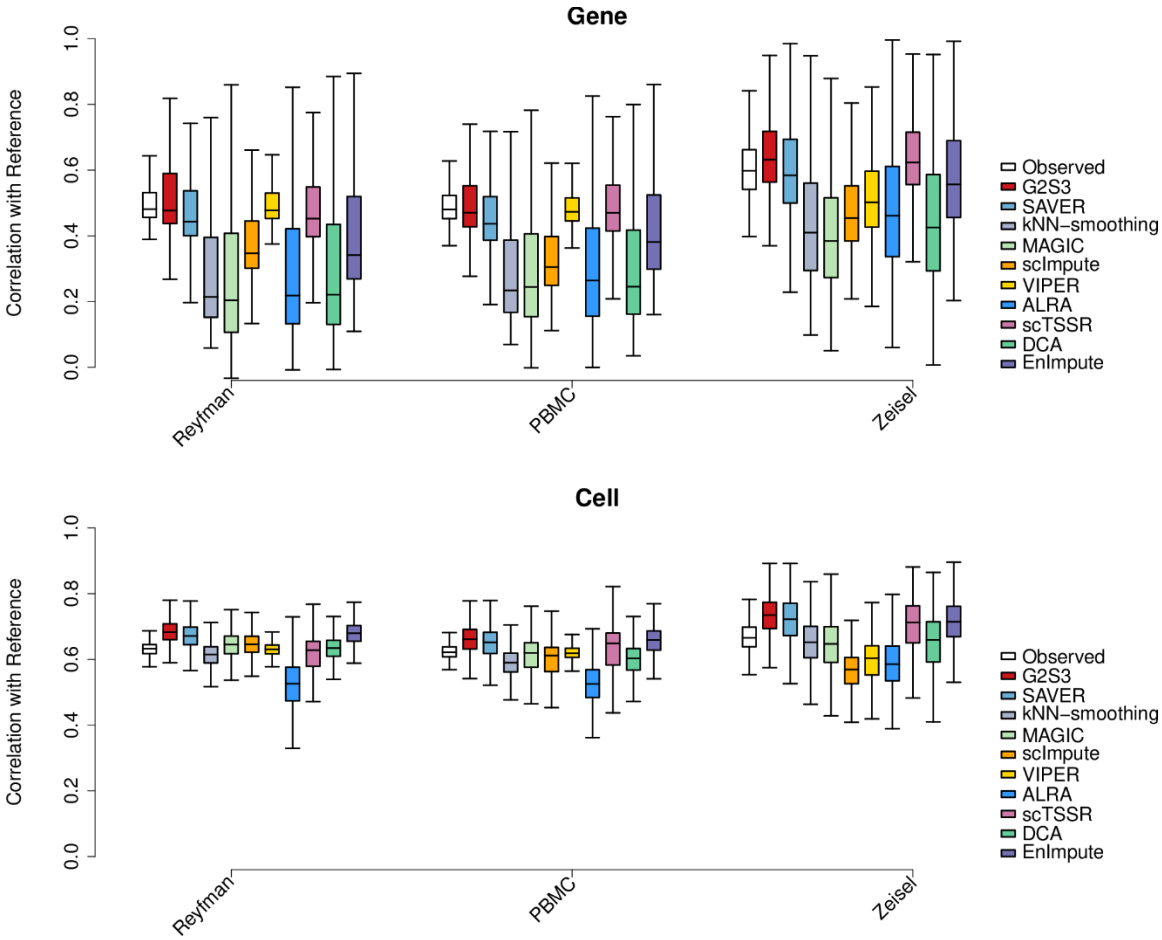


Figure 3.1. Evaluation of expression data recovery of G2S3 by down-sampling. Performance of imputation methods measured by correlation with reference data from the first category of datasets, using gene-wise (top) and cell-wise (bottom) correlation. Box plots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile (whiskers). Outlier data beyond this range are not shown.

3.4.4 Restoration of cell subtype separation

The second category of datasets was used to assess the performance of imputation methods in restoring separation between different cell types. In the Chu dataset, there were 7 cell types including two undifferentiated human ES cell lines (H1 and H9), human foreskin fibroblasts (HF), neuronal progenitor cells (NP), definitive endoderm cells (DE), endothelial cells (EC), and trophoblast-like cells (TB). To quantify the performance in separating these cell subtypes, we calculated the ratio of average inter-subtype distance to average intra-subtype distance using the top K principal components (PCs) of the data before and after imputation, for $K = 1, \dots, 50$. We also calculated the silhouette coefficient that measures how similar cells are to cells from the same cell type compared to other cell types. In Fig 3.2, G2S3 and EnImpute had the highest inter/intra-subtype distance ratio and silhouette coefficient. Both methods performed better than the raw unimputed data, while MAGIC, scImpute, ALRA and DCA performed worse than the raw data. SAUCIE performed the worst. These results suggest that G2S3 greatly improved the separation between different cell types by enhancing the biologically meaningful information in the top PCs. Its performance is comparable to EnImpute that takes advantage over several methods.

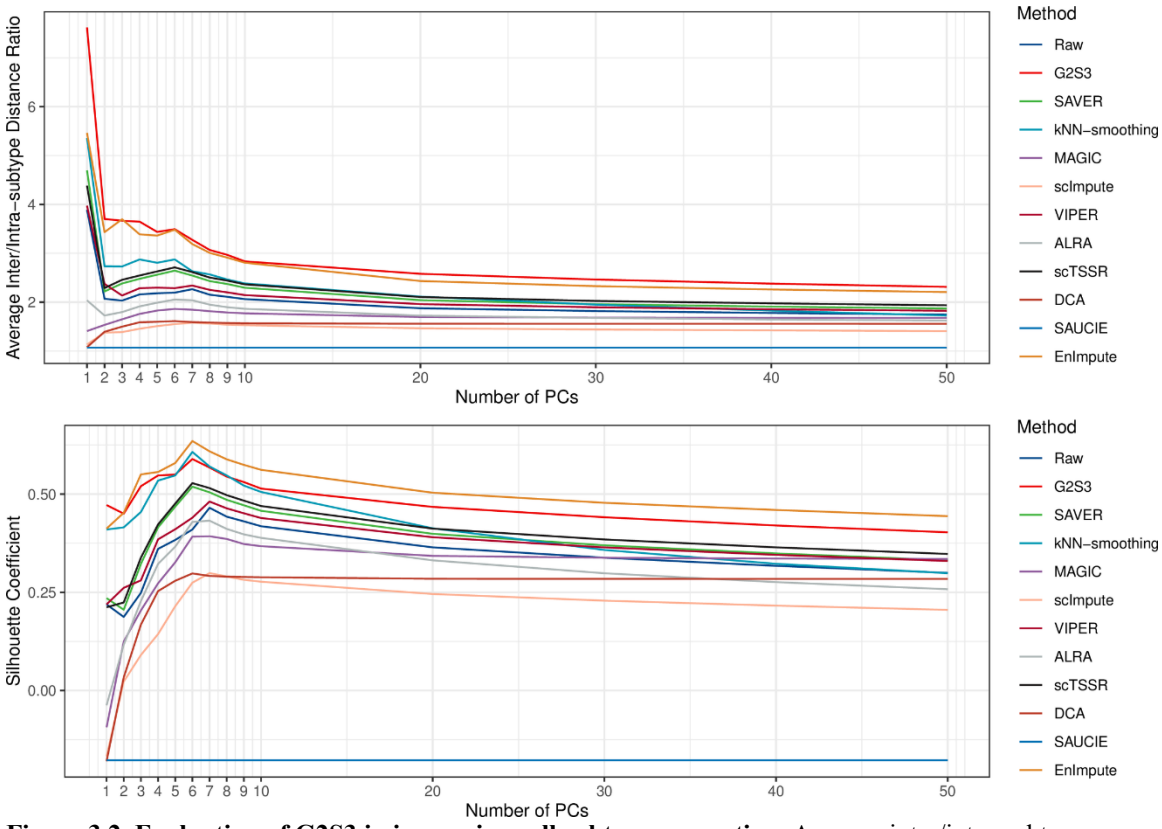


Figure 3.2. Evaluation of G2S3 in improving cell subtype separation. Average inter/intra-subtype distance ratio (top) and silhouette coefficient (bottom) to demonstrate cell subtype separation using the top principal components of the raw unimputed and imputed data by each method in the Chu dataset.

To demonstrate the comparison using cell clustering results, we generated uniform manifold approximation and projection (UMAP) plots in which cells were colored to represent the seven cell types in the original dataset. The normalized mutual information (MI) and adjusted rand index (RI) were calculated to measure the consistency between cell clustering results and true cell subtype labels. Fig 3.3 shows that the imputed data by G2S3 and EnImpute had a better separation of all cell subtypes than the raw unimputed data, except for H1 and H9 cells. Given that both H1 and H9 are undifferentiated human ES cell lines, it is expected that separating them is more difficult due to the relative homogeneity of human ES cells compared to the progenitors. In contrast, the other imputation methods did not have comparable improvement or even reduced the separation of different cell types.

Specifically, DE cells were mixed with EC and TB cells in the raw data and were not separated from the other cell subtypes by all methods except G2S3 and EnImpute. MAGIC was able to separate EC, HF and TB cells from each other and the rest of the cell subtypes, while SAVER was able to separate EC and HF cells from each other and the rest of the cell subtypes. VIPER, ALRA, scTSSR and DCA only separated HF cells from the rest, similar to the raw data. The imputed data by kNN-smoothing formed many small clusters. scImpute tended to mix different cell types into one cluster. SAUCIE overly smoothed the data and was not able to separate any cell types. Based on the two measures of consistency between cell clustering results and true cell subtype labels, EnImpute had the best separation of the cell subtypes (MI=0.77, RI=0.70) and G2S3 was the second best (MI=0.74, RI=0.64), while the other methods did not have comparable performance. Notice that EnImpute is an ensemble method that combines imputation results from multiple methods, and G2S3 is the only method that achieved comparable performance to EnImpute.

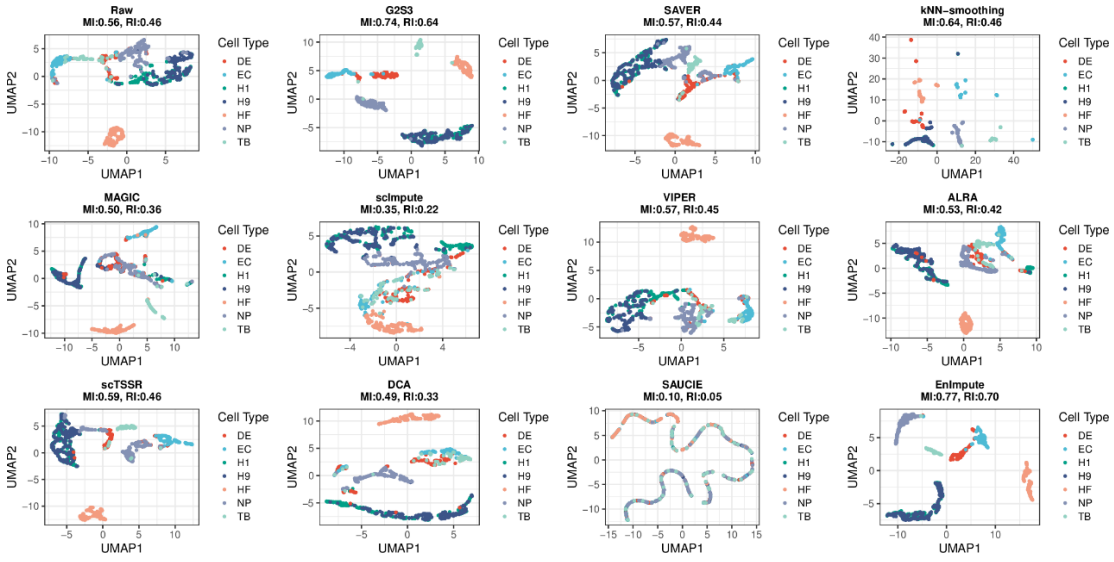


Figure 3.3. Plots showing 2D-Visualization of the Chu dataset. UMAP plots of the raw unimputed and imputed data by all methods. Cells are colored by true cell subtype labels. The normalized mutual

information (MI) and adjusted rand index (RI) are calculated to measure the consistency between cell clustering results and true cell subtype labels.

Fig S3.5 demonstrates the expression of two cell subtype marker genes *GATA*6, a marker gene of DE cells, and *NANOG*, a marker gene of H1/H9 cells [104] across all cells in the raw unimputed and imputed data by all methods. The normalized MI and adjusted RI that measure the consistency between cell clustering results based on these two marker genes and true cell labels for DE and H1/H9 cells were also calculated. We can see that G2S3 provided the best separation between H1/H9 cells, DE cells and other cell subtypes. Specifically, while the raw data mixed H1/H9 cells with other cell subtypes, G2S3 successfully recovered the expression of *GATA*6 and *NANOG* to better separate DE and H1/H9 cell subtypes both from each other and from the other cell subtypes. The cell clustering results on the G2S3 imputed data achieved the highest consistency with true cell subtype labels, indicating its best performance. None of the other methods had comparable performance. DCA separated H1/H9 cells but had DE cells marginally overlapped with other cell types. We observed many small clusters of cells after imputation by kNN-smoothing, similar to the pattern displayed in Fig 3.3. The other methods did not improve cell subtype separation compared to the raw data. In addition, the imputed data by VIPER, kNN-smoothing and ALRA still had a large proportion of dropout events. These results suggest that G2S3 had the best performance in restoring the separation of different cell types, preserving biological meaningful variations, and reducing technical noises.

3.4.5 Improvement in cell trajectory inference

Reconstruction of cell trajectories using scRNA-seq data is important for investigating a dynamic process. However, dropout events may impair pseudo-time inference. We used the Petropoulos dataset to evaluate the performance of all imputation methods in cell trajectory inference. This dataset consists of human preimplantation embryonic cells from five embryonic days (E3-E7) that represent the differentiation stage or age of the embryonic cells. We used Monocle 2 to infer pseudo-time from the raw unimputed and imputed data by each method [110], and compared to the actual embryonic days of the cells for performance evaluation. The pseudotemporal ordering score (POS) and Kendall rank correlation coefficient (Cor) were calculated to measure the consistency. Fig 3.4 shows cell trajectories in the raw and imputed data by all methods. The cell trajectory plots showed the sequential layout of cells from earlier to later embryonic days. The imputed data by G2S3, scImpute, VIPER and EnImpute had the highest consistency with the actual embryonic days, indicating their best performance among all methods. SAVER, kNN-smoothing, MAGIC, ALRA and DCA formed the second tier of methods with lower consistency. scTSSR performed worse than the raw data. SAUCIE had significantly lower consistency (POS=0.07, Cor=0.07) compared to all other methods in cell trajectory inference. Furthermore, the trajectory analysis showed an increased heterogeneity among cells from later embryonic days, especially starting from embryonic day 5. This was consistent with the observation of a significant embryonic cell differentiation event on embryonic day 5 [105].

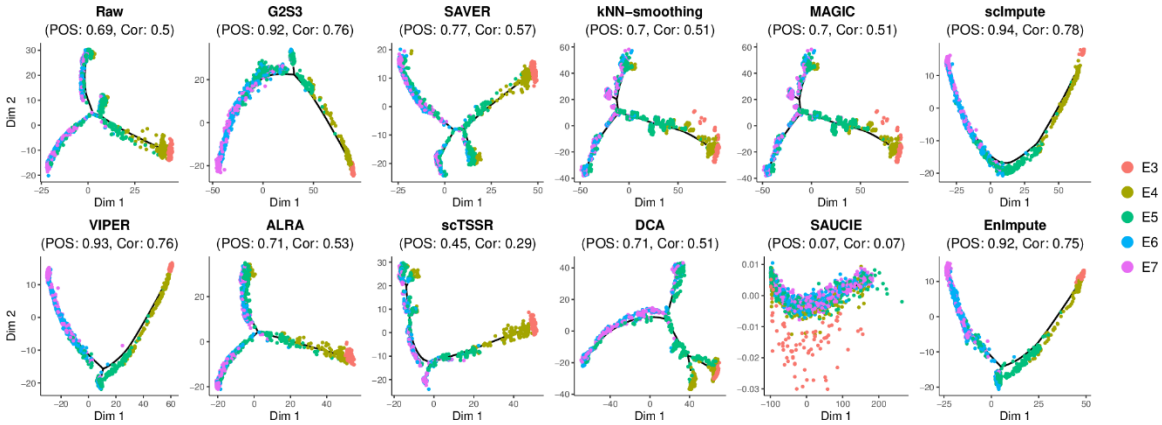


Figure 3.4. Visualization of cell trajectories in the raw and imputed data by all methods. Cells are projected into two-dimensional space using reversed graph embedding. Pseudotemporal ordering score (POS) and Kendall rank correlation coefficient (Cor) are used to measure the consistency between the actual

3.4.6 Improvement in differential expression analysis

One common analytical task for scRNA-seq studies is to identify genes differentially expressed between cells from two groups of subjects or two cell types. In this section, we used two datasets to evaluate and compare the improvement in downstream differential expression analysis before and after imputation by all methods: the Chu dataset of different cell types and the Trapnell dataset of differentiating human myoblasts. Besides the scRNA-seq data, both datasets provide bulk RNA-seq data on the same samples. The differentially expressed genes identified from the bulk RNA-seq data were treated as ground truth. We assessed the predictive power of the scRNA-seq data imputed by different methods on the ground truth using receiver operating characteristic (ROC) curves.

In the Chu dataset, we identified marker genes that differentiate the two cell types: NP and H1 cells. Fig 3.5A shows that G2S3 had the highest area under the curve (AUC) in detecting differentially expressed genes. kNN-smoothing, DCA and EnImpute had an AUC score lower than G2S3 but higher than the raw data. The other methods had comparable performance to the raw data except MAGIC, which had the lowest AUC. This is likely due

to the fact that a small cluster of NP cells were mixed with H1 cells after imputation by MAGIC (Fig 3.3), resulting in compromised performance in marker gene identification. Our results were largely consistent with a previous evaluation of imputation methods in identifying differentially expressed genes using Fluidigm C1 data [113]. No genes achieved significance in the imputed data by SAUCIE so the result of SAUCIE could not be shown. In the Trapnell dataset, we performed differential expression analysis between undifferentiated primary human myoblasts and mature myotubes captured 72 hours after inducing differentiation. Fig 3.5B shows that G2S3 achieved the highest AUC indicating its best performance, followed by VIPER. kNN-smoothing and DCA had much worse performance than the raw data. No genes achieved significance in the imputed data by MAGIC and SAUCIE so their results could not be shown. Altogether, the results from both datasets showed that G2S3 had the best improvement in the downstream differential expression analysis.

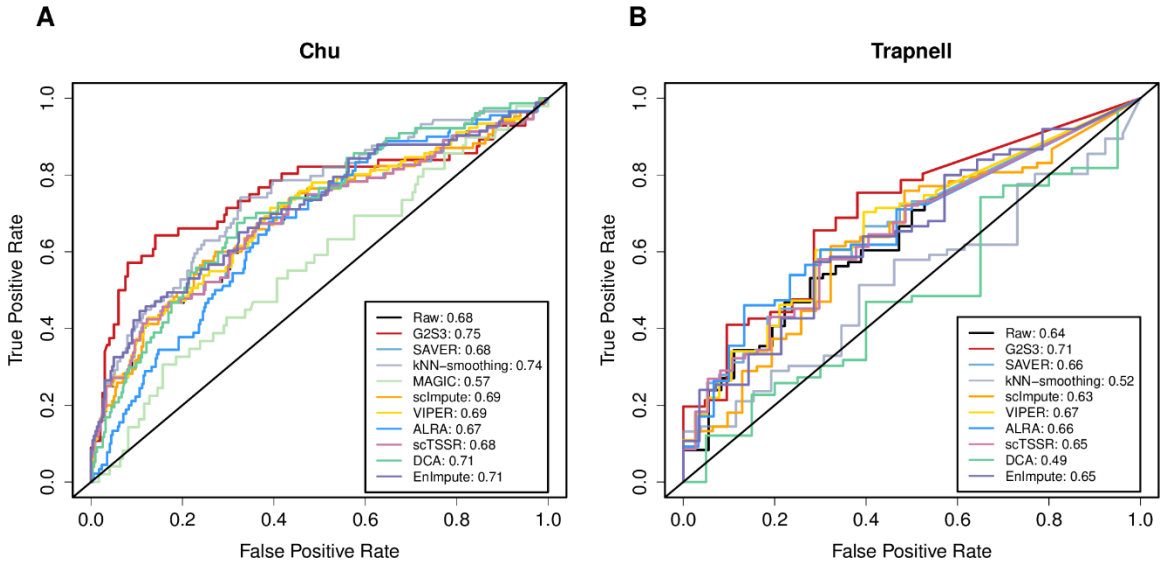


Figure 3.5. Receiver operating characteristic (ROC) curves demonstrating improvement in differential expression analysis. ROC curves of the scRNA-seq differential expression results predicting differentially expressed genes identified in the bulk RNA-seq data on the same samples in the Chu (A) and Trapnell (B) datasets.

3.4.7 Gene correlation relationship recovery

We compared the method performance in recovering gene correlation relationships using two scRNA-seq dastasets. In the Paul dataset, we examined the pairwise correlation between well-known transcription factors in the development of blood cells before and after imputation [111]. In the Buettner dataset, we investigated the relationships among a set of 67 periodic marker genes before and after imputation, in which 16 genes have peak expression in the G1/S phase and 51 genes have peak expression in the G2/M phase [108].

In the Paul dataset, the regulatory relationship among key regulators of the transcriptional differentiation of megakaryocyte/erythrocyte progenitors and granulocyte/macrophage progenitors in the raw data and the imputed data by each method were used for performance evaluation. The gene regulatory network (GRN) of these regulators was established in a previous study based on biological experiments [114–116] and served as the ground truth. We reconstructed GRNs by two methods, GENIE3 [117] and PPCOR [118], in the raw and imputed datasets. The inferred GRNs were compared to the ground-truth network using the area under the precision-recall curve (AUPRC). For each imputation method, we reported the AUPRC ratio (AUPRC divided by that of a random predictor) with 50 replications. Fig 3.6 demonstrates that G2S3 achieved the highest AUPRC ratio, followed by kNN-smoothing, using both GRN inference methods. The AUPRC ratios of GRNs inferred from the imputed data by either MAGIC or SAUCIE were much lower than that from a random predictor, suggesting that the gene regulatory relationships were distorted after imputation.

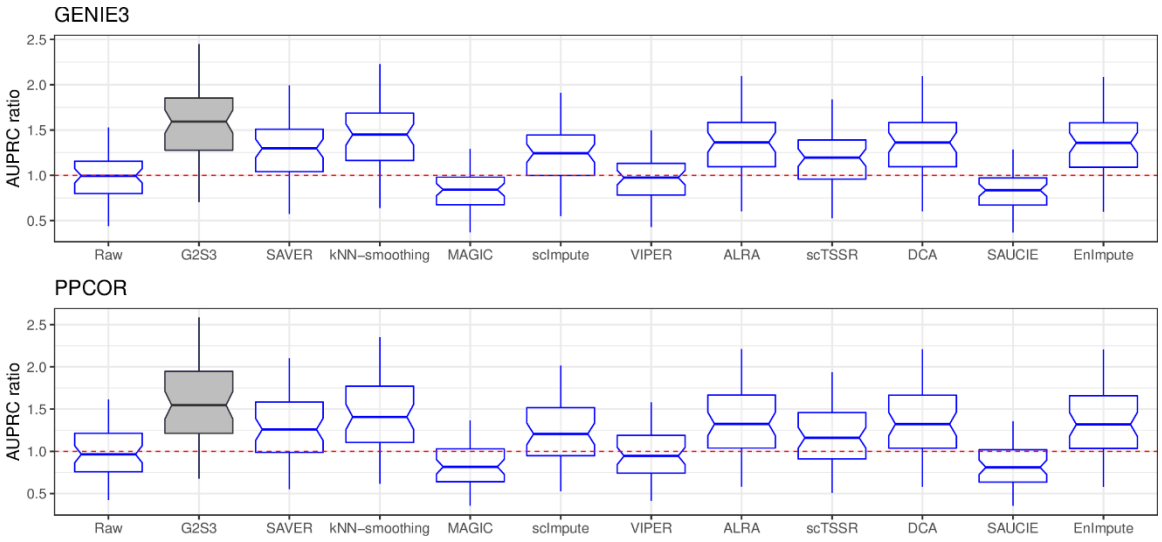


Figure 3.6. Performance of G2S3 in recovering gene regulatory relationship. Boxplots showing the area under the precision-recall curve (AUPRC) ratios that measure the accuracy of inferred GRNs using the imputed data by different imputation methods. Both GENIE3 (top) and PPCOR (bottom) were used to infer GRNs. Red line indicates the performance of a random predictor.

We also examined the pairwise correlations between these key regulators. Based on previous studies [114–116], inhibitory and activatory gene pairs were defined, among which inhibitory pairs were expected to have negative correlation while activatory pairs were expected to have positive correlation. The mutually inhibitory pairs of genes include *Fli*1 vs. *Klf*1, *Egr*1 vs. *Gfi*1, *Cebpa* vs. *Gata*1, and *Sfpi*1 vs. *Gata*1; and the mutually activatory pairs include *Sfpi*1 vs. *Cebpa*, *Zfpm*1 vs. *Gata*1, *Klf*1 vs. *Gata*1. Fig S3.6 shows that most of the methods were able to enhance the pairwise correlations after imputation in the correct direction. Overall, G2S3 and SAVER showed the greatest enhancement of pairwise correlation for both inhibitory and activatory pairs, followed by kNN-smoothing and EnImpute. Although MAGIC intensified the pairwise correlations, most activatory pairs had correlations close to 1 after imputation. ALRA and DCA strengthened the

pairwise correlation for activatory pairs but did not improve much for inhibitory pairs. Imputation by SAUCIE resulted in all gene pairs to be highly positively correlated.

In the Buettner dataset, we expect pairs of periodic genes whose expression peak in the same phase of cell cycle to be positively correlated and those that peak during different phases to be negatively correlated. There are 67 marker genes for G1/S and G2/M phases [111]. We examined the correlation of all 2,211 marker gene pairs in the raw data and imputed data by each method. The proportion of gene pairs whose correlations are in the correct direction was used for performance comparison. Table 3.2 shows that all methods had comparable performance in maintaining a high proportion of positively correlated gene pairs, whereas their performance varies in restoring negatively correlated gene pairs. G2S3, SAVER and EnImpute were able to recover 28% or more of the negatively correlated gene pairs. All gene pairs became positively correlated after imputation by MAGIC, scImpute, VIPER, ALRA, DCA and SAUCIE, thus no negative correlation was observed after imputation. Similar observations were found in a previous study in which some of these methods introduced a large number of positive gene correlations after imputation, many of which may be spurious [90].

Table 3.2. Fraction of periodic gene pairs with correct direction of correlation in the raw and imputed data by each method

Imputation Methods	Positive Pairs	Negative Pairs
Raw	1.00	0.00
G2S3	0.91	0.32
SAVER	0.94	0.28
kNN-smoothing	0.97	0.17
MAGIC	1.00	0.00
scImpute	1.00	0.00
VIPER	1.00	0.00
ALRA	1.00	0.00

scTSSR	0.98	0.11
DCA	1.00	0.00
SAUCIE	1.00	0.00
EnImpute	0.91	0.46

In summary, the results from both datasets suggested that G2S3 enhanced gene-gene relationships especially for negatively correlated gene pairs in which the expression of one gene is inhibited by the other. As lowly expressed genes are in general harder to impute, negatively correlated relationship is a harder task for imputation to restore the correlated relationship.

3.4.8 Summary of method performance

We evaluated and compared the performance of G2S3 and the other ten imputation methods using five evaluation criteria corresponding to five downstream analyses of scRNA-seq data. Fig 3.7 summarizes the overall performance of all methods. G2S3 was ranked first in three out of the five evaluation criteria, second in cell clustering, and third in cell trajectory inference. For those criteria under which G2S3 did not achieve the best performance, it had close or comparable performance to the best method. No other method achieved the best performance in as many criteria as that of G2S3. Overall, G2S3 performed the best among all the methods, followed by EnImpute and VIPER.

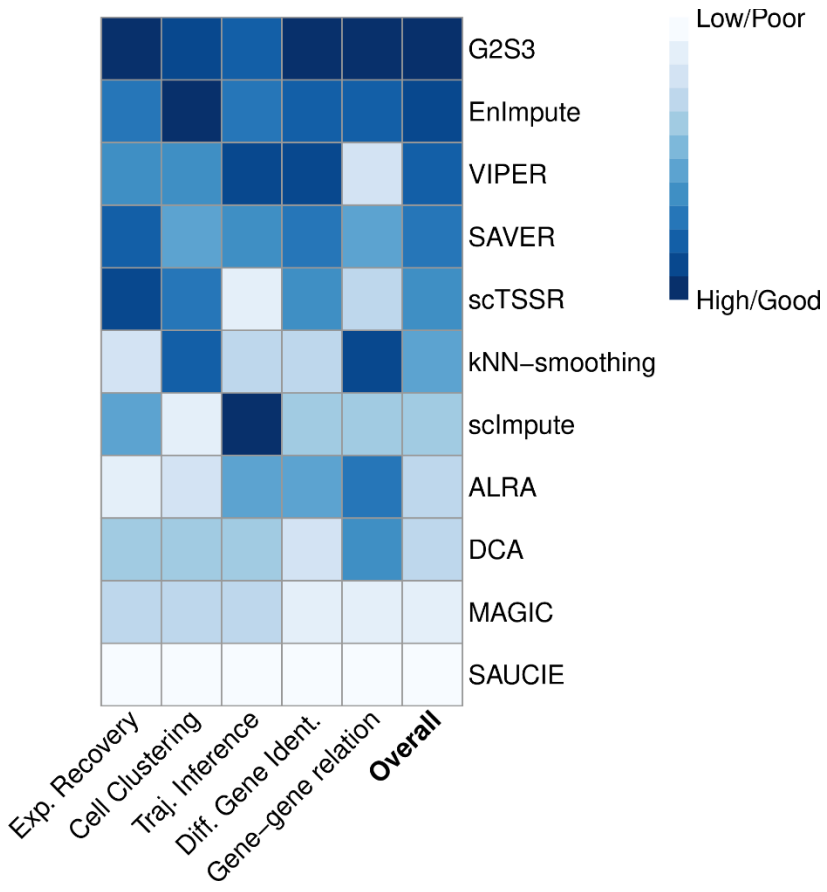


Figure 3.7. Summary of performance of G2S3 and other imputation methods. A heatmap demonstrating the method performance based on the five evaluation criteria. The left five columns display performance rank using each of the five evaluation criteria. The rightmost column displays the overall performance rank based on the sum of all five ranks.

3.4.9 Computation time

While SAVER and EnImpute have comparable performance to G2S3 in some datasets, G2S3 is computationally more efficient (Table 3.3). Since both G2S3 and SAVER are gene network-based imputation methods, their computation time is expected to increase with the number of genes to be imputed. This makes gene network-based methods more suitable than those based on cell similarity for large scRNA-seq datasets with tens or even hundreds of thousands of cells. In real data analysis, G2S3 was on average about twenty times faster

than SAVER. EnImpute is an ensemble method that relies on imputation results from multiple methods and therefore takes longer time than SAVER. On the other hand, the computation time of imputation methods that borrow information from similar cells increases dramatically with the number of cells in the data. As demonstrated in a previous study, scImpute and VIPER were unable to scale beyond 10K cells within 24 hours [95]. In our assessment, VIPER takes about two days to impute the down-sampled datasets with several thousands of genes while other methods finish within several minutes.

Table 3.3. Computational Time for Each Imputation Methods. Running time in minutes for each imputation task among imputation methods using a single processor on an 8-core, 50 GB RAM, Intel Xeon 2.6 GHz CPU machine. *Derived computing time sum over five methods and ensemble computing time.

	G2S3	SAVER	kNN-smoothing	MAGIC	scImpute	VIPER	ALRA	scTSSR	DCA	SAUCIE	EnImpute*
Reyfman	4.27	60.12	0.25	0.35	29.46	5289.17	0.16	9.80	5.40	0.86	102.23
Zeisel	2.99	43.26	0.18	0.24	70.67	3618.86	0.10	4.26	4.27	0.74	121.84
PBMC	1.09	25.91	0.15	0.17	17.77	524.37	0.08	3.48	2.78	0.98	50.84

3.5 Discussion

We have developed a novel method G2S3 to impute dropouts in scRNA-seq data. G2S3 learns a sparse and smooth signals of gene graph from scRNA-seq data and borrows information from nearby genes in the graph for imputation. We evaluated and compared the performance of G2S3 and ten existing imputation methods in terms of recovering true expression levels, restoring cell subtype separation, reconstructing cell trajectory, identifying differentially expressed genes, and restoring gene correlation relationships using eight scRNA-seq datasets. The results demonstrated that G2S3 achieved superior performance or had comparable performance to other methods based on the five evaluation

criteria above, especially for genes with relatively low expression. Furthermore, G2S3 is the most computationally efficient method for large-scale scRNA-seq data imputation.

Unlike imputation methods that borrow information across similar cells, G2S3 harnesses the structural relationship among genes obtained through graph signal processing to perform imputation. Using eight real datasets, we showed that methods relying on cell similarity tend to remove biological variation among cells and intensify subject-level batch effects. In contrast, G2S3 enhances cell subtype separation and thus relatively reduces variations in cells from the same cell type and subject. The down-sampling and differential expression analysis results showed that G2S3 outperformed the other methods, especially for lowly expressed genes. Of note, imputation methods such as SAVER, scImpute and VIPER, used parametric models for gene expression. However, as the noise distribution varies across different scRNA-seq platforms, assumptions of the parametric models may be violated, particularly for new technologies. Graph signal processing extracts signals from data by optimizing a smoothness regulated objective function, so it is in principle less sensitive to the noise distribution. To our knowledge, there are two imputation methods that use gene graph/network for imputation in scRNA-seq data, published during the preparation of this manuscript: netNMF-sc [119] uses network-regularized non-negative matrix factorization to leverage gene-gene interactions for imputation; and netSmooth [120] incorporates protein-protein interaction networks to smooth gene expression values. Both methods require prior information on gene-gene interactions from RNA-seq or microarray studies of bulk tissue. In contrast, G2S3 learns gene network structure in an unbiased way from scRNA-seq data. In our experiments, G2S3 had comparable performance to EnImpute, an ensemble learning method that combines results from multiple imputation methods.

G2S3 learns gene-gene relationship by optimizing a sparse gene graph and at the same time allows expression levels to change smoothly between closely connected genes. Since many gene networks and biochemical networks are sparse [110,121,122], the sparsity property is important for inferring gene network. There are several methods available for constructing gene network, many of them are kernel-based, which result in full weight matrices where sparsity has to be imposed afterwards, for example, thresholding the adjacency weights. We found that top eigenvectors of graph Laplacian on the gene networks learned from Gaussian kernel were highly correlated with dropout rate, suggesting that dropout events tend to bias the construction of gene network in scRNA-seq data. G2S3 algorithm uses one step random walk to avoid over-smoothing because multiple steps of the random walk tended to overly smooth data and leaded to worse performance. Similar observations were reported in another manuscript discussing parameter tuning for diffusion-based imputation methods for scRNA-seq data [112]. It showed that for many diffusion-based methods including MAGIC, single step (t=1) yielded better performance than multiple steps or iterations until convergence. For UMI-based datasets, to account for the effect of varying sequencing depths, we recommend normalizing UMI counts before applying G2S3 for more accurate construction of gene graph and imputation of expression levels.

Despite the advantages of G2S3 over the other imputation methods shown in this article, G2S3 can be improved in several directions. First, G2S3 uses a lazy random walk on the gene graph to recover dropout events, i.e., weighted average of the observed expression of the gene of interest and that from neighboring genes. The weights currently depend only on between gene similarity which can be improved by considering the reliability of

observed read counts, cell library size, and dispersion of gene expression, similar to the weights used in SAVER. Second, G2S3 does not consider dropout rate and therefore imputes all values at once. This can be improved by calculating the probability of being a dropout for each observed read count and only performing imputation on those with a high dropout probability. Third, the G2S3 model can be improved by adding two tuning parameters for the second and third terms in the objective function that control the degree of smoothness and sparsity of the resulting gene network. The tuning parameters can be chosen based on the complexity and structure of scRNA-seq data. Finally, G2S3 does not consider the potential subject effect in the data, which has been shown to be prevalent and dominant in certain cell types. One way to address this issue is to consider subject effect as "batch" effect and remove it using batch effect removal tools. This is effective only when there are no other effects of interest confounding the subject effect, for example, disease effect, because they will also be removed together with "batch" effect. When there are other effects that confound with subject effect and are the interest of study, G2S3 can be improved to consider subject effect and disease effect at the same time in imputation.

Supplementary Materials

Supplementary methods

S 2.1 Quasi-likelihood score of $\boldsymbol{\gamma}$

The quasi-likelihood score function for $\boldsymbol{\gamma}$ can be written as

$$\begin{aligned}\boldsymbol{U}(\boldsymbol{\gamma}) &= (\mathbf{I}_m \otimes \boldsymbol{\Phi}\boldsymbol{A})^T(\boldsymbol{\Sigma}_G \otimes \boldsymbol{\Phi})^{-1}\left(\widetilde{\boldsymbol{G}} - 2\boldsymbol{p} \otimes \mathbf{1}_n - \boldsymbol{\gamma} \otimes \boldsymbol{\Phi}\boldsymbol{A}\right) \\ &= (\boldsymbol{\Sigma}_G^{-1} \otimes \boldsymbol{A})^T\left(\widetilde{\boldsymbol{G}} - 2\boldsymbol{p} \otimes \mathbf{1}_n - \boldsymbol{\gamma} \otimes \boldsymbol{\Phi}\boldsymbol{A}\right) \\ &= \boldsymbol{\Sigma}_G^{-1}\boldsymbol{G}^T\boldsymbol{A} - (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\boldsymbol{\Sigma}_G^{-1}\boldsymbol{\gamma}.\end{aligned}$$

The last equation holds because $\boldsymbol{A}^T\mathbf{1}_n = 0$ and $(\boldsymbol{\Sigma}_G^{-1} \otimes \boldsymbol{A})^T\widetilde{\boldsymbol{G}} = \boldsymbol{\Sigma}_G^{-1}\boldsymbol{G}^T\boldsymbol{A}$.

S 2.2 Covariance of $\boldsymbol{G}^T\boldsymbol{A}$

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{G}^T\boldsymbol{A}) &= \mathrm{Cov}\left((\mathbf{I}_m \otimes \boldsymbol{A})^T\widetilde{\boldsymbol{G}}\right) = (\mathbf{I}_m \otimes \boldsymbol{A})^T(\boldsymbol{\Sigma}_G \otimes \boldsymbol{\Phi})(\mathbf{I}_m \otimes \boldsymbol{A}) \\ &= \boldsymbol{\Sigma}_G \otimes (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A}) = (\boldsymbol{A}^T\boldsymbol{\Phi}\boldsymbol{A})\boldsymbol{\Sigma}_G\end{aligned}$$

Supplementary Figures

Figure S3.1 Comparison of the mean-variance relationship in gene expression before and after down-sampling. For each gene, the coefficient of variation (CV) across all cells after down-sampling (y-axis) is plotted against the CV of non-zero cells in the reference data (x-axis).

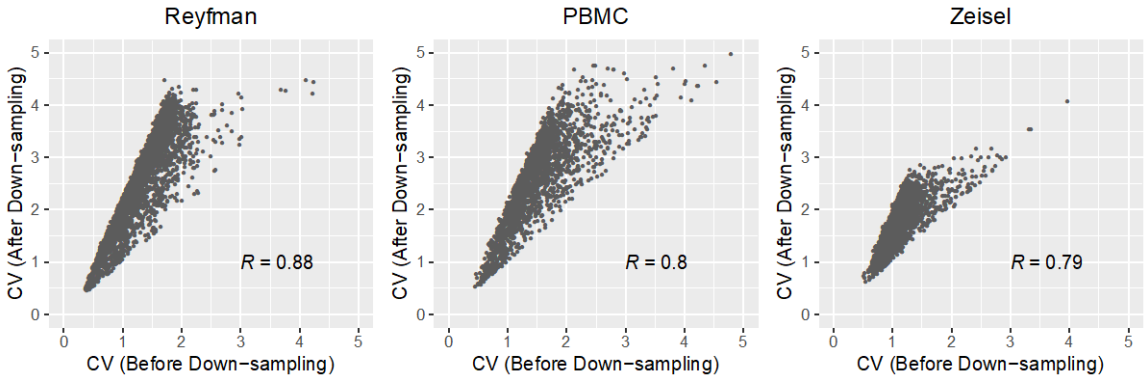


Figure S3.2 Optimal value of hyperparameter in G2S3. A. Mean squared error (MSE) at different diffusion steps in three down-sampled datasets. B. Gene-wise and cell-wise correlations of G2S3 imputed data at different diffusion steps and the reference data.

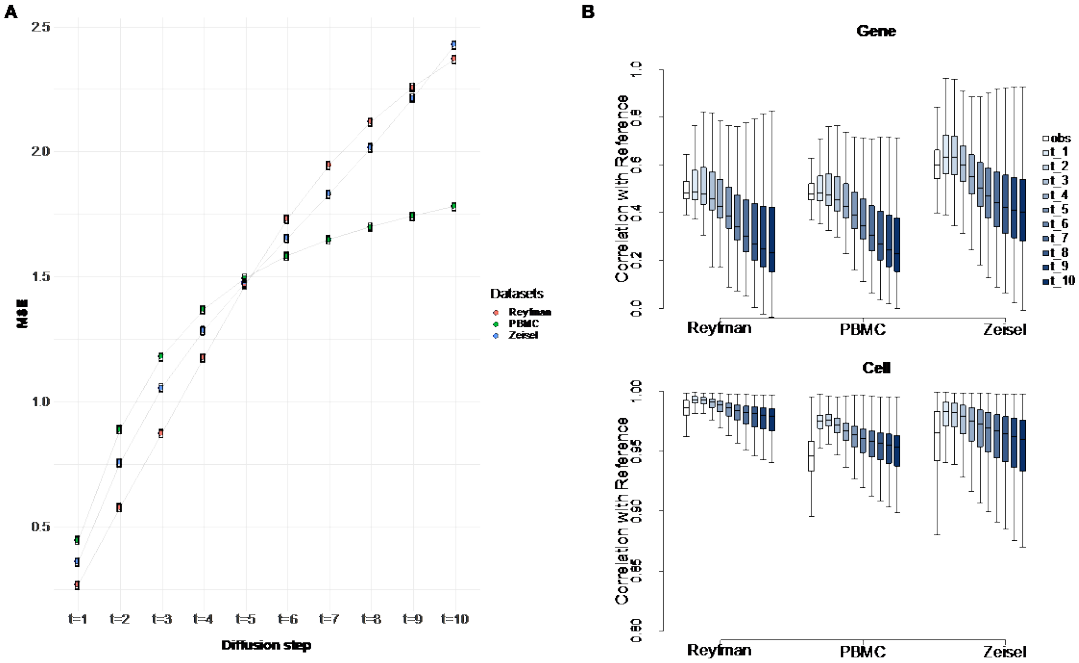


Figure S3.3 Evaluation of expression data recovery of all imputation methods by down-sampling. Performance of imputation methods measured by correlation with reference data from the first category of datasets, using gene-wise (top) and cell-wise (bottom) correlation. Box plots show the median (center line), interquartile range (hinges), and 1.5 times the interquartile (whiskers). Outlier data beyond this range are not shown.

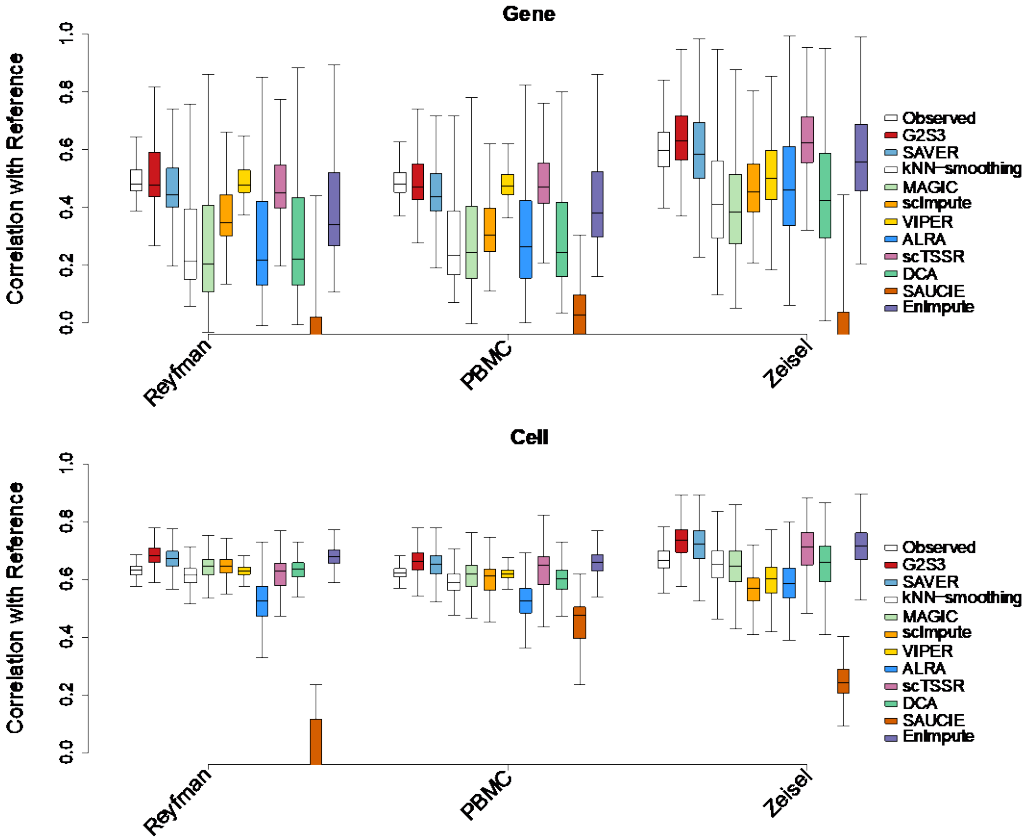


Figure S3.4 Evaluation of expression data recovery of all imputation methods by down-sampling in three gene strata. Performance of imputation methods measured by correlation with reference data from the first category of datasets, using gene-wise (top) and cell-wise (bottom) correlation. Genes are stratified into three groups: widely (>80%, left), mildly (30%-80%, middle), and rarely (<30%, right) expressed.

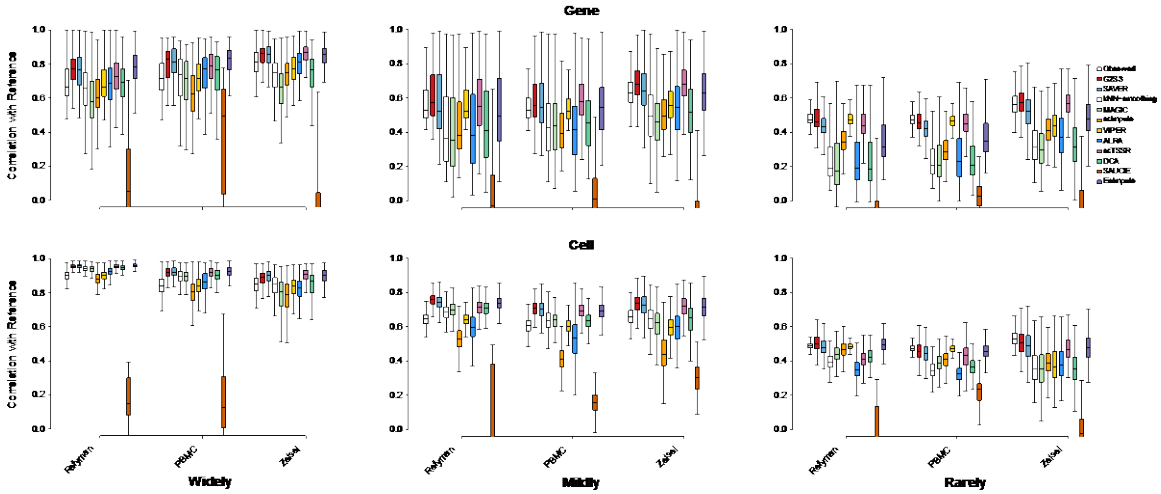


Figure S3.5 Cell subtype marker gene expression in the Chu dataset. Scatter plot showing expression level of marker genes for DE cells (*GATA6*) and H1/H9 cells (*NANOG*). Cells are colored by the cell subtype labels.

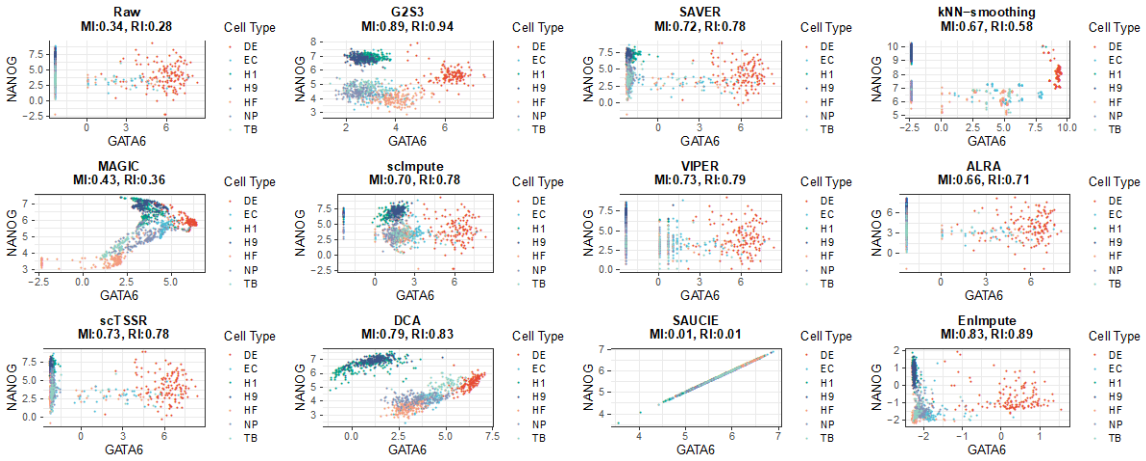
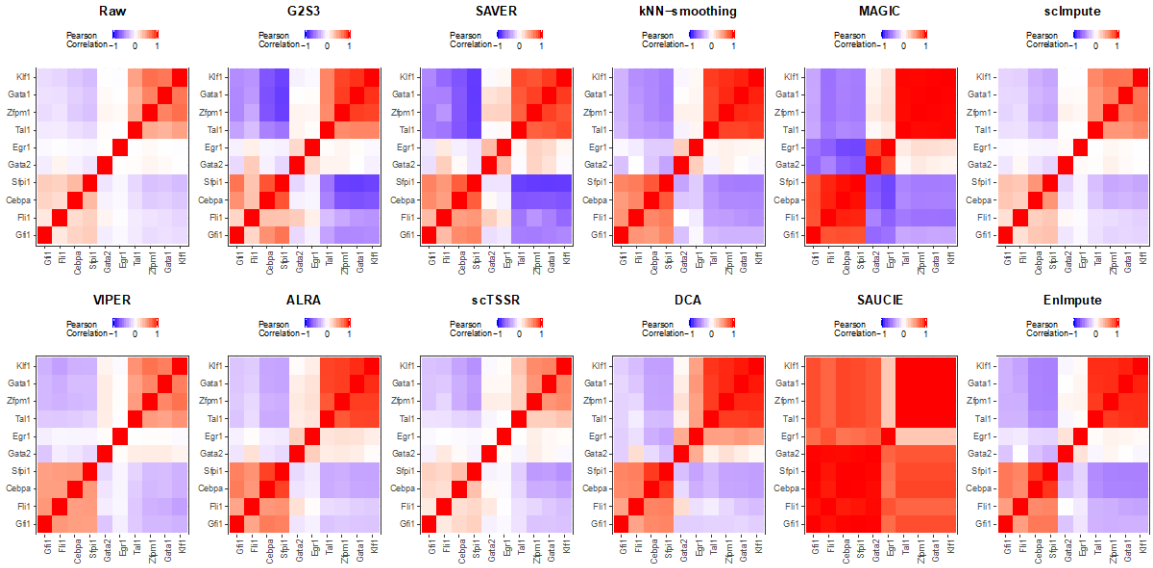


Figure S3.6 Evaluation of recovering gene correlation relationship of all imputation methods in the Paul dataset. Heatmaps of pairwise correlations between well-known blood regulators.



Bibliography

1. Furlotte NA, Eskin E, Eyheramendy S. Genome-Wide Association Mapping With Longitudinal Data. Genetic Epidemiology. 2012;36(5):463–71.

2. Sikorska K, Rivadeneira F, Groenen PJF, Hofman A, Uitterlinden AG, Eilers PHC, et al. Fast linear mixed model computations for genome-wide association studies with longitudinal data. Statistics in Medicine. 2013 Jan;32(1):165–80.

3. Sitlani CM, Rice KM, Lumley T, Mcknight B, Cupples LA, Avery CL, et al. Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. Statistics in Medicine. 2015 Jan;34(1):118–30.

4. Das K, Li J, Wang Z, Tong C, Fu G, Li Y, et al. A dynamic model for genome-wide association studies. Human Genetics. 2011 Jun;129(6):629–39.

5. Londono D, Chen KM, Musolf A, Wang R, Shen T, Brandon J, et al. A novel method for analyzing genetic association with longitudinal phenotypes. Statistical Applications in Genetics and Molecular Biology. 2013 Jan;12(2):241–61.

6. Meirelles OD, Ding J, Tanaka T, Sanna S, Yang HT, Dudekula DB, et al. SHAVE: Shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. European Journal of Human Genetics. 2013 Jun;21(6):673–9.

7. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics. 2018 Sep;50(9):1335–41.

8. Jiang D, Mbatchou J, McPeek MS. Retrospective Association Analysis of Binary Traits: Overcoming Some Limitations of the Additive Polygenic Model. Human Heredity. 2015;80(4):187–95.

9. Hayeck TJ, Zaitlen NA, Loh PR, Vilhjalmsson B, Pollack S, Gusev A, et al. Mixed model with correction for case-control ascertainment increases association power. American Journal of Human Genetics. 2015 May;96(5):720–30.

10. Jiang D, Zhong S, McPeek MS. Retrospective Binary-Trait Association Test Elucidates Genetic Architecture of Crohn Disease. The American Journal of Human Genetics. 2016 Feb 4;98(2):243–55.

11. Jakobsdottir J, McPeek MS. MASTOR: Mixed-Model Association Mapping of Quantitative Traits in Samples with Related Individuals. Am J Hum Genet. 2013 May 2;92(5):652–66.

12. Zhong S, Jiang D, McPeek MS. CERAMIC: Case-Control Association Testing in Samples with Related Individuals, Based on Retrospective Mixed Model Analysis with Adjustment for Covariates. PLoS Genetics. 2016 Oct;12(10):e1006329.

13. Hayeck TJ, Loh PR, Pollack S, Gusev A, Patterson N, Zaitlen NA, et al. Mixed model association with family-biased case-control ascertainment. American Journal of Human Genetics. 2017 Jan;100(1):31–9.

14. Wu X, McPeek MS. L-GATOR: Genetic Association Testing for a Longitudinally Measured Quantitative Trait in Samples with Related Individuals. The American Journal of Human Genetics. 2018 Apr 5;102(4):574–91.

15. Schildcrout JS, Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. Biostatistics. 2008 Oct;9(4):735–49.

16. Schildcrout JS, Schisterman EF, Aldrich MC, Rathouz PJ. Outcome-related, auxiliary variable sampling designs for longitudinal binary data. Epidemiology. 2018 Jan;29(1):58–66.

17. Schildcrout JS, Schisterman EF, Mercaldo ND, Rathouz PJ, Heagerty PJ. Extending the case-control design to longitudinal data: stratified sampling based on repeated binary outcomes. Epidemiology. 2018 Jan;29(1):67–75.

18. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population structure and relatedness for binary traits in genetic association studies via logistic mixed models. American Journal of Human Genetics. 2016 Apr;98(4):653–66.

19. Wang Z, Xu K, Zhang X, Wu X, Wang Z. Longitudinal SNP-set association analysis of quantitative phenotypes. Genetic Epidemiology. 2017 Jan;41(1):81–93.

20. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993 Mar;88(421):9–25.

21. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Research. 2005 Nov;15(11):1576–83.

22. Justice AC, Dombrowski E, Conigliaro J, Fultz SL, Gibson D, Madenwald T, et al. Veterans Aging Cohort Study (VACS): Overview and description. Medical care. 2006 Aug;44(8 Suppl 2):S13-24.

23. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics. 2009 Jun;5(6):e1000529.

24. Gelernter J, Sherva R, Koesterer R, Almasy L, Zhao H, Kranzler HR, et al. Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. Molecular Psychiatry. 2014 Jun;19(6):717–23.

25. Pierucci-Lagha A, Gelernter J, Feinn R, Cubells JF, Pearson D, Pollastri A, et al. Diagnostic reliability of the semi-structured assessment for drug dependence and alcoholism (SSADDA). Drug and Alcohol Dependence. 2005 Dec;80(3):303–12.

26. Muthén B. Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data. In: The SAGE Handbook of Quantitative Methodology for the Social Sciences [Internet]. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc.; p. 346–69. Available from: http://methods.sagepub.com/book/the-sage-handbook-of-quantitative-methodology-for-the-social-sciences/n19.xml

27. Jung J, Tawa EA, Muench C, Rosen AD, Rickels K, Lohoff FW. Genome-wide association study of treatment response to venlafaxine XR in generalized anxiety disorder. Psychiatry Research. 2017 Aug;254:8–11.

28. Wong ML, Arcos-Burgos M, Liu S, Vélez JI, Yu C, Baune BT, et al. The PHF21B gene is associated with major depression and modulates the stress response. Molecular Psychiatry. 2017 Jul;22(7):1015–25.

29. Wang KS, Liu XF, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophrenia Research. 2010 Dec;124(1–3):192–9.

30. Goes FS, Mcgrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics. 2015 Dec;168(8):649–59.

31. Li Z, Chen J, Yu H, He L, Xu Y, Zhang D, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nature Genetics. 2017 Nov;49(11):1576–83.

32. Li Q, Wineinger NE, Fu DJ, Libiger O, Alphs L, Savitz A, et al. Genome-wide association study of paliperidone efficacy. Pharmacogenetics and Genomics. 2017 Jan;27(1):7–18.

33. Ribeiro EA, Scarpa JR, Garamszegi SP, Kasarskis A, Mash DC, Nestler EJ. Gene network dysregulation in dorsolateral prefrontal cortex neurons of humans with cocaine use disorder. Scientific Reports. 2017 Dec;7(1):5412.

34. Zhou B, Shi J, Whittemore AS. Optimal methods for meta-analysis of genome-wide association studies. Genetic Epidemiology. 2011 Nov;35(7):581–91.

35. Le Merrer J, Becker JAJ, Befort K, Kieffer BL. Reward processing by the opioid system in the brain. Physiological Reviews. 2009 Oct;89(4):1379–412.

36. Soderman AR, Unterwald EM. Cocaine-induced mu opioid receptor occupancy within the striatum is mediated by dopamine D2 receptors. Brain Research. 2009 Nov;1296:63–71.

37. Bahi A, Dreyer JL. Cocaine-induced expression changes of axon guidance molecules in the adult rat brain. Molecular and Cellular Neuroscience. 2005 Feb;28(2):275–91.

38. Jassen AK. Receptor regulation of gene expression of axon guidance molecules: implications for adaptation. Molecular Pharmacology. 2006 Jul;70(1):71–7.

39. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013 Jun;45(6):580–5.

40. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204–13.

41. Hyman SE, Malenka RC. Addiction and the brain: The neurobiology of compulsion and its persistence. Nature Reviews Neuroscience. 2001 Oct;2(10):695–703.

42. Warlow SM, Robinson MJF, Berridge KC. Optogenetic central amygdala stimulation intensifies and narrows motivation for cocaine. The Journal of Neuroscience. 2017 Aug;37(35):3141–16.

43. Fan R, Zhang Y, Albert PS, Liu A, Wang Y, Xiong M. Longitudinal Association Analysis of Quantitative Traits. Genetic Epidemiology. 2012 Sep;36(8):856–69.

44. Wang Z, Xu K, Zhang X, Wu X, Wang Z. Longitudinal SNP-set association analysis of quantitative phenotypes. Genetic Epidemiology. 2017 Jan;41(1):81–93.

45. Wang Z, Wang N, Wu R, Wang Z. fGWAS: An R package for genome-wide association analysis with longitudinal phenotypes. Journal of Genetics and Genomics. 2018 Jul;45(7):411–3.

46. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013 Jun 6;92(6):841–53.

47. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012 Sep;13(4):762–75.

48. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008 Sep;83(3):311–21.

49. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011 Jul 15;89(1):82–93.

50. Sun J, Zheng Y, Hsu L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. Genetic Epidemiology. 2013;37(4):334–44.

51. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014 Aug;197(4):1081–95.

52. Chen H, Huffman JE, Brody JA, Wang C, Lee S, Li Z, et al. Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am J Hum Genet. 2019 Feb 7;104(2):260–74.

53. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. The American Journal of Human Genetics. 2019 Mar;104(3):410–21.

54. He Z, Lee S, Zhang M, Smith JA, Guo X, Palmas W, et al. Rare-variant association tests in longitudinal studies, with an application to the Multi-Ethnic Study of Atherosclerosis (MESA). Genet Epidemiol. 2017;41(8):801–10.

55. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLOS Genetics. 2009 Feb 13;5(2):e1000384.

56. Wu W, Wang Z, Xu K, Zhang X, Amei A, Gelernter J, et al. Retrospective Association Analysis of Longitudinal Binary Traits Identifies Important Loci and Pathways in Cocaine Use. Genetics [Internet]. 2019 Oct 7 [cited 2019 Dec 3]; Available from: https://www.genetics.org/content/early/2019/10/07/genetics.119.302598

57. Goetz MB, Madenwald T, Gibert CL, Leaf DA, Rimland D, Oursler KAK, et al. Veterans Aging Cohort Study (VACS). Medical Care. 2006 Aug;44(Suppl 2):S13–24.

58. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.

59. de Vries PS, Brown MR, Bentley AR, Sung YJ, Winkler TW, Ntalla I, et al. Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. Am J Epidemiol. 2019 01;188(6):1033–54.

60. Melendez RI, McGinty JF, Kalivas PW, Becker HC. Brain region-specific gene expression changes after chronic intermittent ethanol exposure and early withdrawal in C57BL/6J mice. Addiction Biology. 2012;17(2):351–64.

61. Winham SJ, Cuellar-Barboza AB, Oliveros A, McElroy SL, Crow S, Colby C, et al. Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. Mol Psychiatry. 2014 Sep;19(9):1010–6.

62. Strakowski SM, DelBello MP. The co-occurrence of bipolar and substance use disorders. Clinical Psychology Review. 2000 Mar 1;20(2):191–206.

63. Virkkunen M, Goldman D, Linnoila M. Serotonin in alcoholic violent offenders. Ciba Found Symp. 1996;194:168–77; discussion 177-182.

64. Seo D, Patrick CJ, Kennealy PJ. Role of Serotonin and Dopamine System Interactions in the Neurobiology of Impulsive Aggression and its Comorbidity with other Clinical Disorders. Aggress Violent Behav. 2008 Oct;13(5):383–95.

65. Lejuez CW, Magidson JF, Mitchell SH, Sinha R, Stevens MC, Wit HD. Behavioral and Biological Indicators of Impulsivity in the Development of Alcohol Use, Problems, and Disorders. Alcoholism: Clinical and Experimental Research. 2010;34(8):1334–45.

66. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22(3):343–52.

67. Kimmey BA, Ostroumov A, Dani JA. 5-HT2A receptor activation normalizes stress-induced dysregulation of GABAergic signaling in the ventral tegmental area. PNAS. 2019 Dec 26;116(52):27028–34.

68. Gould TD, Manji HK. The Wnt Signaling Pathway in Bipolar Disorder. Neuroscientist. 2002 Oct 1;8(5):497–511.

69. Hoseth EZ, Krull F, Dieset I, Mørch RH, Hope S, Gardsjord ES, et al. Exploring the Wnt signaling pathway in schizophrenia and bipolar disorder. Transl Psychiatry. 2018 Mar 6;8(1):1–10.

70. Hack SP, Christie MJ. Adaptations in adenosine signaling in drug dependence: therapeutic implications. Crit Rev Neurobiol. 2003;15(3–4):235–74.

71. Ruby CL, Adams C, Knight EJ, Nam HW, Choi D-S. An Essential Role for Adenosine Signaling in Alcohol Abuse. Curr Drug Abuse Rev. 2010 Sep;3(3):163–74.

72. O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, et al. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. Genome Biol [Internet]. 2018 Nov 12

[cited 2020 Feb 28];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6231252/

73. Peoples LL. Will, Anterior Cingulate Cortex, and Addiction. Science. 2002 May 31;296(5573):1623–4.

74. Mashhoon Y, Czerkawski C, Crowley DJ, Cohen-Gilbert JE, Sneider JT, Silveri MM. Binge alcohol consumption in emerging adults: anterior cingulate cortical 'thinness' is associated with alcohol use patterns. Alcohol Clin Exp Res. 2014 Jul;38(7):1955–64.

75. Zahr NM, Pitel A-L, Chanraud S, Sullivan EV. Contributions of Studies on Alcohol Use Disorders to Understanding Cerebellar Function. Neuropsychol Rev. 2010 Sep 1;20(3):280–9.

76. Kaplan JS, Nipper MA, Richardson BD, Jensen J, Helms M, Finn DA, et al. Pharmacologically Counteracting a Phenotypic Difference in Cerebellar GABAA Receptor Response to Alcohol Prevents Excessive Alcohol Consumption in a High Alcohol-Consuming Rodent Genotype. J Neurosci. 2016 Aug 31;36(35):9019–25.

77. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015 Sep;525(7568):251–5.

78. Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, et al. Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids De Novo to Contribute to Immune Homeostasis. Cell Reports. 2014 May 22;7(4):1130–42.

79. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015 Jan;18(1):145–53.

80. Frishberg A, Peshes-Yaloz N, Cohn O, Rosentul D, Steuerman Y, Valadarsky L, et al. Cell composition analysis of bulk genomics using single-cell data. Nat Methods. 2019 Apr;16(4):327–32.

81. Wu YE, Pan L, Zuo Y, Li X, Hong W. Detecting Activated Cell Populations Using Single-Cell RNA-Seq. Neuron. 2017 Oct 11;96(2):313-329.e6.

82. Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, Guo G, et al. Challenges and emerging directions in single-cell analysis. Genome Biology. 2017 May 8;18(1):84.

83. Shalek AK, Benson M. Single-cell analyses to tailor treatments. Science Translational Medicine [Internet]. 2017 Sep 20 [cited 2020 Apr 7];9(408). Available from: https://stm.sciencemag.org/content/9/408/eaan4730

84. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014 Jul;11(7):740–2.

85. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. bioRxiv. 2018 Apr 9;217737.

86. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174(3):716-729.e27.

87. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature Communications. 2018;9(1):1–9.

88. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018 Dec 8;19(1):220.

89. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biology. 2018 Dec 12;19(1):196.

90. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: Gene expression recovery for single-cell RNA sequencing. Nature Methods. 2018;15(7):539–42.

91. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. bioRxiv. 2018 Aug 22;397588.

92. Jin K, Ou-Yang L, Zhao X-M, Yan H, Zhang X-F. scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation. Bioinformatics. 2020 May 1;36(10):3131–8.

93. Talwar D, Mongia A, Sengupta D, Majumdar A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. Scientific Reports. 2018 Dec;8(1):16329.

94. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications. 2019 Dec;10(1):390.

95. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biology. 2019 Oct 18;20(1):211.

96. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, et al. Exploring single-cell data with deep multitasking neural networks. Nature Methods. 2019 Nov;16(11):1139–45.

97. Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Res [Internet]. 2019 Mar 5 [cited 2020 Apr 7];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6415334/

98. Zhang X-F, Ou-Yang L, Yang S, Zhao X-M, Hu X, Yan H. EnImpute: imputing dropout events in single-cell RNA-sequencing data via ensemble learning. Bioinformatics. 2019 Nov 1;35(22):4827–9.

99. Kalofolias V. How to Learn a Graph from Smooth Signals. In: Artificial Intelligence and Statistics [Internet]. 2016 [cited 2020 Mar 21]. p. 920–9. Available from: http://proceedings.mlr.press/v51/kalofolias16.html

100. Komodakis N, Pesquet J-C. Playing with Duality: An Overview of Recent Primal-Dual Approaches for Solving Large-Scale Optimization Problems. 2014 Jun 20;

101. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, et al. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. American Journal of Respiratory and Critical Care Medicine. 2018 Dec;rccm.201712-2410OC.

102. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Jan 16;8(1):1–12.

103. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, Manno GL, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 Mar 6;347(6226):1138–42.

104. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biology. 2016 Aug 17;17(1):173.

105. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016 May;165(4):1012–26.

106. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014 Apr;32(4):381–6.

107. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell. 2015 Dec 17;163(7):1663–77.

108. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing

data reveals hidden subpopulations of cells. Nature Biotechnology. 2015 Feb;33(2):155–60.

109. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987 Nov 1;20:53–65.

110. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nature Methods. 2017 Oct;14(10):979–82.

111. Dominguez D, Tsai Y-H, Gomez N, Jha DK, Davis I, Wang Z. A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. Cell Research. 2016 Aug;26(8):946–62.

112. Tjärnberg A, Mahmood O, Jackson CA, Saldi G-A, Cho K, Christiaen LA, et al. Optimal tuning of weighted kNN- and diffusion-based methods for denoising single cell genomics data. PLOS Computational Biology. 2021 Jan 7;17(1):e1008569.

113. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biology. 2020 Aug 27;21(1):218.

114. Krumsiek J, Marr C, Schroeder T, Theis FJ. Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. Pesce M, editor. PLoS ONE. 2011 Aug 10;6(8):e22649.

115. Rekhtman N, Radparvar F, Evans T, Skoultchi AI. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. Genes Dev. 1999 Jun 1;13(11):1398–411.

116. Iwasaki H, Mizuno S, Wells RA, Cantor AB, Watanabe S, Akashi K. GATA-1 Converts Lymphoid and Myelomonocytic Progenitors into the Megakaryocyte/Erythrocyte Lineages. Immunity. 2003 Sep 1;19(3):451–62.

117. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010 Sep 28;5(9).

118. Kim S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun Stat Appl Methods. 2015 Nov;22(6):665–74.

119. Elyanow R, Dumitrascu B, Engelhardt BE, Raphael BJ. netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. Genome Res. 2020 Jan 28;gr.251603.119.

120. Ronen J, Akalin A. netSmooth: Network-smoothing based imputation for single cell RNA-seq. F1000Research. 2018;7(0):8.

121. Cai X, Bazerque JA, Giannakis GB. Inference of Gene Regulatory Networks with Sparse Structural Equation Models Exploiting Genetic Perturbations. PLOS Computational Biology. 2013 May 23;9(5):e1003068.

122. Kikkawa A. Random Matrix Analysis for Gene Interaction Networks in Cancer Cells. Scientific Reports. 2018 Jul 13;8(1):10607.