Multi-SNP mediation intersection-union test

Wujuan Zhong[1], Cassandra N. Spracklen[2], Karen L. Mohlke[2], Xiaojing Zheng[3,*], Jason Fine[1,4,*] and Yun Li[1,2,5,*]

[1]Department of Biostatistics, [2]Department of Genetics, [3]Department of Pediatrics, [4]Department of Statistics and Operations Research and [5]Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 19, 2018; revised on March 19, 2019; editorial decision on April 14, 2019; accepted on April 16, 2019

Abstract

Summary: Tens of thousands of reproducibly identified GWAS (Genome-Wide Association Studies) variants, with the vast majority falling in non-coding regions resulting in no eventual protein products, call urgently for mechanistic interpretations. Although numerous methods exist, there are few, if any methods, for simultaneously testing the mediation effects of multiple correlated SNPs via some mediator (e.g. the expression of a gene in the neighborhood) on phenotypic outcome. We propose multi-SNP mediation intersection-union test (SMUT) to fill in this methodological gap. Our extensive simulations demonstrate the validity of SMUT as well as substantial, up to 92%, power gains over alternative methods. In addition, SMUT confirmed known mediators in a real dataset of Finns for plasma adiponectin level, which were missed by many alternative methods. We believe SMUT will become a useful tool to generate mechanistic hypotheses underlying GWAS variants, facilitating functional follow-up.

Availability and implementation: The R package SMUT is publicly available from CRAN at https://CRAN.R-project.org/package=SMUT.

Contact: xiaojinz@email.unc.edu or jfine@email.unc.edu or yunli@med.unc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWASs) have been successful for detecting genetic variants associated with complex diseases and traits. The effects of genetic variants either individually or even in aggregation on complex traits are typically small to moderate at best. More importantly, the vast majority of GWAS variants reside in non-coding regions, with largely elusive underlying mechanism. Expression quantitative trait loci (eQTL) analysis (Lloyd-Jones *et al.*, 2017; The GTEx Consortium, 2015; Yang *et al.*, 2017) has facilitated functional interpretation. Transcriptome-wide association studies (TWAS), which identify association between imputed gene expression and trait, also generates mechanistic hypotheses (Gamazon *et al.*, 2015; Gusev *et al.*, 2016)

(BioRxiv: https://doi.org/10.1101/286013). Such integration of genotype, gene expression and phenotype information from GWAS and eQTL datasets will fundamentally advance our knowledge of molecular mechanisms of complex disorders and traits. Several excellent review papers exist for causal relationship inference in the context of genetic mapping for complex traits (Ainsworth *et al.*, 2017; Civelek and Lusis, 2014).

Integrative genomic studies enable mechanistic interpretations, e.g. via either the methods of instrumental variable(s) [IV(s)] and/or mediation analysis. Mendelian randomization (MR) framework (Lawlor *et al.*, 2008; Smith and Ebrahim, 2003) treats genetic variant(s) as the IV(s) to assess causal effects of genetic variants through some mediator(s) of interest [e.g. expression levels of some gene(s)] on the trait of interest. Classic MR methods make several key assumptions including complete mediation, where single nucleotide polymorphisms (SNPs) must be marginally independent of the confounding between mediator and outcome, and a priori knowledge that the causal flow is from SNP to mediator but not the reverse (Lawlor *et al.*, 2008). Violating the assumptions leads to invalid IV

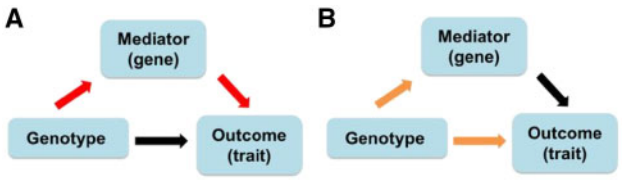

Fig. 1. Directed acyclic graph for mediation and pleiotropy. (**A**) Red arrows indicate the mediation effect of the genotype on the outcome through the mediator. (**B**) Orange arrows indicate the pleiotropy

analysis and biased inference. Some more recent MR methods allow relaxation of certain key assumption(s). Such relaxation(s), however, are at costs. For example, MR-Egger (Barfield *et al.*, 2018; Bowden *et al.*, 2015) relaxes the complete mediation assumption and allows multiple IVs/SNPs by first analyzing each IV individually and then meta-analyzing individual IV results. However, MR-Egger assumes that these multiple SNPs are uncorrelated, limiting its application to a typical locus where multiple partially dependent SNPs exist.

Another drawback of MR methods is that they cannot distinguish between mediation and pleiotropy, the phenomenon of one SNP effecting on multiple outcomes (at either molecular or organism level) (Fig. 1). Since pleiotropy is commonly observed (Solovieff *et al.*, 2013), MR methods are not preferred, when the goal is to infer mediation or to generate mechanistic hypotheses. The recent CaMMEL method (BioRxiv: https://doi.org/10.1101/219428), extending the MR-Egger framework, allows correlated IVs and incomplete mediation beyond the multiple mediators modeled and, to distinguish mediation from pleiotropy. CaMMEL, designed for multiple mediators modeled simultaneously, is sub-optimal for single mediator analysis. In addition, CaMMEL assumes the presence of at least one eQTL since it tests the effect of mediators on phenotype, in contrast to testing SNP(s) effect via mediator(s) on phenotype. Similar to CaMMEL, MR-BMA (BioRxiv: https://doi.org/10.1101/396333), also adopts a Bayesian framework, leverages summary statistics, and accommodates multiple mediators and multiple IVs. Unlike CaMMEL, MR-BMA assumes uncorrelated IVs and complete mediation through the modeled mediators, and aims primarily at selecting true mediators from a set of correlated mediators.

Besides TWAS and MR, other mechanism elucidating methods in the recent literature include causal inference test (CIT) (Millstein *et al.*, 2009) and Huang *et al.* (Huang, 2015; Huang *et al.*, 2014). CIT employs a regression-based framework to test for complete mediation, which is a largely unrealistic scenario. The methods of Huang *et al.* adopt a kernel regression framework and uses variance component score statistic (Lin, 1997) to test for mediation. However, these methods also assume that genetic variants tested contain *a priori* known eQTL(s), and similar to CaMMEL, test the effect of mediators on phenotype, in contrast to testing indirect SNP(s) effect through mediators on phenotype.

Popular classic mediation approaches include causal steps, difference method and product method (MacKinnon *et al.*, 2007; VanderWeele, 2016). The causal steps approach performs multiple tests involved in a causal chain. The difference method is based on the difference in the coefficient estimate of the treatment (here, a SNP) before and after including the mediator in the regression model. The product method, such as the Sobel test (Sobel, 1982), explicitly tests the product of the treatment coefficient in the mediator model and the mediator coefficient in the outcome model. These methods, commonly adopted to test the mediation effect of a single genetic variant on a trait through a single mediator, have been evaluated by Barfield *et al.* (2017). However, it is unclear that such methods can be adapted to integrative genomic settings with high dimensional SNPs.

In short, to the best of our knowledge, few, if any, existing methods can simultaneously accommodate incomplete mediation as well as multiple correlated SNPs, when complete individual level data (including genotype, mediator and phenotype information) are available. To fill the gap, we propose here multi-SNP mediation intersection-union test (SMUT) to explicitly accommodate both direct and indirect (via mediator) effect of multiple (in the order of hundreds to thousands) correlated SNPs on phenotype of interest. SMUT is a flexible, regression-based approach that evaluates the joint mediation effects of multiple genetic variants on some trait of interest through a single mediator. SMUT extends the classic framework of Baron and Kenny (1986) to allow multiple treatment variables (in our context, multiple genetic variants). Leveraging the intersection-union test (IUT) (Berger and Hsu, 1996), SMUT decomposes mediation effects using two separate regression models. One is the mediator model where we regress the mediator on multiple genetic variants. For this mediator model, SMUT adopts the SKAT (Lee *et al.*, 2014; Wu *et al.*, 2011) framework to handle a potentially large number of genetic variants in a statistically and computationally efficient manner. The other is the outcome model where we regress the outcome on both the mediator and multiple genetic variants. Classic regression models fail for the outcome model due to the high dimensionality of the SNPs. To solve this issue in SMUT's outcome model, we adopt a mixed effects model and the Rao's score test (Engle, 1984; Radhakrishna Rao and Bartlett, 1948) for mediation testing. Our extensive simulations and real data analysis demonstrate the advantages of SMUT over alternative methods. For example, with controlled Type-I error, we show up to 92% power gain in simulations. More importantly, in real data analysis, SMUT confirms mediations at several well-established positive control loci while most of the alternative methods failed to reveal any of the relationships.

2 Materials and methods

2.1 SMUT

SMUT is a powerful test for the joint mediation effects of multiple genetic variants on a trait through a single mediator. The multiple genetic variants can be in a region, sub-locus defined by genes, or moving windows across the genome.

2.2 Notation and data set-up

Without loss of generality, we assume that we have three types of data. Specifically, genotypes, gene expression measurements (can be other types of mediators such as metabolite levels or protein abundances) and phenotypic trait are available. Let $G = (G_1, G_2, \ldots, G_q)$ be the n by q genotype matrix, where n is sample size, q is the total number of marker and $G_j = (G_{1j}, G_{2j}, \ldots, G_{nj})^T$ is the vector of genotypes for the n samples at marker j, $j = 1, 2, \ldots, q$. We consider an additive model with G_{ij} taking values 0, 1, 2, measuring the number of copies of the minor allele. Suppose in total there are l genes $M, M^{(2)}, M^{(3)}, \ldots, M^{(l)}$, with the first notation M having no superscript. Here, $M = (M_1, M_2, \ldots, M_n)^T$, is the vector of expression values of a given gene (the mediator) for n samples. Similarly, $M^{(2)}, \ldots, M^{(l)}$ are the vectors of expression values of the other $(l-1)$ genes (i.e. mediators). Let $Y = (Y_1, Y_2, \ldots, Y_n)^T$ be the vector of phenotypic trait.

2.3 SMUT model and test for joint mediation effects

SMUT models the effects of genetic variants on the trait mediated by the expression level of a single gene. We start with considering a

more general model with multiple genes expression levels via the following regression models:

$$Y = \alpha_1 + M\theta + \sum_{k=2}^{l} M^{(k)}\theta^{(k)} + G\gamma + \epsilon_1 \quad (1)$$

$$\begin{cases} M = \alpha_2 + G\beta + \epsilon_2 \\ M^{(2)} = \alpha^{(2)} + G\beta^{(2)} + \epsilon^{(2)} \\ M^{(3)} = \alpha^{(3)} + G\beta^{(3)} + \epsilon^{(3)} \\ \cdots \\ M^{(l)} = \alpha^{(l)} + G\beta^{(l)} + \epsilon^{(l)} \end{cases} \quad (2)$$

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^T$ are the direct effects of the q genetic variants; $\beta\theta$ measures the indirect effects mediated by M for the multiple genetic variants. Similarly $\beta^{(k)}\theta^{(k)}$ measures the indirect effects mediated by $M^{(k)}$, $k = 2, 3, \ldots, l$.
Substituting the $M, M^{(2)}, M^{(3)}, \ldots, M^{(l)}$ with the values in (2), we have

$$Y = \tilde{\alpha} + G\beta\theta + G\tilde{\gamma} + \tilde{\epsilon} \quad (3)$$

where $\tilde{\alpha} = \alpha_1 + \alpha_2\theta + \sum_{k=2}^{l} \alpha^{(k)}\theta^{(k)}$, $\tilde{\gamma} = \gamma + \sum_{k=2}^{l} \beta^{(k)}\theta^{(k)}$, $\tilde{\epsilon} = \epsilon_1 + \theta\epsilon_2 + \sum_{k=2}^{l} \theta^{(k)}\epsilon^{(k)}$ Equation (3) shows that indirect effects mediated by $M^{(2)}, M^{(3)}, \ldots, M^{(l)}$ would be absorbed by the direct effects $\tilde{\gamma}$ if we only model gene M. Therefore, without loss of generality, we only consider the mediation analysis for a given single gene expression level and consider the regression models below

$$Y = \alpha_1 + M\theta + G\gamma + \epsilon_1 \quad \text{Outcome model} \quad (4)$$

$$M = \alpha_2 + G\beta + \epsilon_2 \quad \text{Mediator model} \quad (5)$$

where $\epsilon_1 \sim N(0, \sigma_1^2 I)$, $\epsilon_2 \sim N(0, \sigma_2^2 I)$, and we assume that ϵ_1 and ϵ_2 are independent; otherwise their correlation would make themselves mediator-outcome confounders which violates the key assumption for mediation analysis (MacKinnon *et al.*, 2007; VanderWeele, 2016).

Here γ measures effects from two sources: direct effects of the q genetic variants on outcome; and indirect effects of genetic variants via mediators other than M. For presentation brevity and clarity, we hereafter use direct effects to refer to the aggregated effects from the above two sources. We are interested in testing the mediation effect, of the q genetic variants via mediator M. Specifically, we test the null hypothesis $H_0 : \beta\theta = 0$, which is divided into two sub-hypotheses, $H_0^\theta : \theta = 0$ and $H_0^\beta : \beta = 0$. This then can be conveniently solved by the IUT (Supplementary Section 1).

2.4 Testing β in the mediator model and θ in the outcome model

Many of the testing methods for association between multiple genetic variants and the trait can be applied here. We adopt the SKAT framework, a de facto locally powerful test (Ionita-Laza *et al.*, 2013; Wu *et al.*, 2010, 2011), which efficiently accommodates large numbers of genetic variants, including both common and rare variants.

The outcome model is also high dimensional with multiple genetic effects and the mediator. Classic regression models tend to fail for such models. As a solution, we employ the following mixed effects model to reduce the dimension of parameters.

$$\begin{cases} \gamma_j \sim_{i.i.d.} N(\mu_\gamma, \sigma_\gamma^2) \\ \epsilon_i \sim_{i.i.d.} N(0, \sigma_\epsilon^2) \\ Y_i | (\gamma_1, \ldots, \gamma_q, G) = \alpha_1 + M_i\theta + \Sigma_{j=1}^{q} G_{ij}\gamma_j + \epsilon_i \end{cases} . \quad (6)$$

We adopt an Expectation–maximization algorithm to obtain maximum likelihood estimate under the null hypothesis and derive a score test statistic for the fixed effect θ in the presence of a high dimensional SNPs, modeled as random effects (Supplementary Section 2).

2.5 Simulations

To evaluate the performance of SMUT in comparison with alternative methods, we carried out extensive simulations to investigate power and Type-I error. We first simulated 20 000 European-like chromosomes in a 1 Mb region, using the COSI coalescent model (Schaffner *et al.*, 2005) to generate realistic data in terms of allele frequency, linkage disequilibrium and population differentiation. The final dataset had 23 889 SNPs in a 1 Mb region. We constructed 10 000 pseudo-individuals by pairing up the 20 000 simulated chromosomes. To evaluate power and Type-I error, we generated 200 datasets with 1000 samples each by sampling without replacement from the entire pool of 10 000 samples above. Simulations were restricted to the 2891 SNPs with minor allele frequency (MAF) $\geq$1%.

The outcome (trait) and the mediator were generated via the following outcome model (7) and mediator model (8), respectively.

$$Y = \alpha_1 + M\theta + (sSNPs\ and\ oSNPs)\gamma + \epsilon_1 \quad (7)$$

$$M = \alpha_2 + (sSNPs\ and\ mSNPs)\beta + \epsilon_2. \quad (8)$$

Where $\epsilon_1 \sim N(0, 1)$, $\epsilon_2 \sim N(0, 1)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_q)^T, \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^T$, $\beta_j \sim_{i.i.d.} c_\beta N(2, 2)$, $\gamma_j \sim_{i.i.d.} c_\gamma N(2, 2)$, $j = 1, 2, \ldots, q$.

We set $c_\gamma = 0.2$ to evaluate the performance of SMUT and alternative methods under the scenario of pleiotropy. Specifically, the shared SNPs (sSNPs) between the two models are those that influence both the mediator and the outcome trait. The outcome (or mediator) specific SNPs only contribute to the trait (or mediator). The causal SNPs are the union of the sSNPs, mediator specific SNPs (mSNPs) and outcome specific SNPs (oSNPs). We considered two scenarios in terms of causal SNP density: sparse and dense (Table 1), with 10 and 1000 causal SNPs, respectively. The set of (10 or 1000) causal SNPs, common across the 200 datasets, were randomly selected from the 2891 SNPs with MAF $\geq$ 1%. β and γ, again fixed across the 200 datasets, were independently drawn from a normal distribution with mean and variance both being 2. Error terms ϵ_1 and ϵ_2 were independently generated from standard normal distribution and were separately simulated for each of the 200 datasets.

In the simulations, we tested the joint mediation effects of these 2891 SNPs on the trait using SMUT and other methods including adapted Huang *et al.*'s method, adapted LASSO (Tibshirani, 1996), adapted CaMMEL, Sobel test and SMR (Zhu *et al.*, 2016). The Huang *et al.*'s method only tests mediator effect in the outcome model, assuming *a priori* the presence of SNPs' effects on mediator (i.e. non-zero β), adopting a kernel framework where effect(s) of interest are treated as random and SNPs as fixed (Huang, 2015;

Table 1. Causal SNP composition in two simulated scenarios

	No. of causal SNPs	No. of sSNPs	No. of mSNPs	No. of oSNPs
Sparse	10	4	3	3
Dense	1000	334	333	333

Note: The sparse(dense) scenario is to simulate datasets based on a small(large) number of causal SNPs. Causal SNPs are the union of sSNPs, mSNPs and oSNPs. sSNPs have effects on both mediator and outcome. Mediator(outcome) specific SNPs have effects only on mediator(outcome). All these SNPs are randomly selected from the 2891 SNPs with MAF$\geq$1%.

Huang *et al.*, 2014), in contrast to our outcome model where SNPs are treated as random and mediator of interest as fixed. For fair comparisons across methods, i.e. testing both β and θ, we applied the original Huang *et al.* for the outcome model and SKAT for the mediator model, then combined tests from the two models via IUT, integrating the variance component score test in the outcome model (from the original Huang *et al.*) and score test from SKAT in the mediator model. The adapted LASSO employs LASSO for variable selection in the outcome model and applies IUT using regular regression with the selected variables in the outcome model and all the variables (i.e. genetic variants) via SKAT framework in the mediator model. The adapted CaMMEL applies IUT on CaMMEL results to test θ in the outcome model and SKAT result to test β in the mediator model. Since Sobel test and SMR can only model one single SNP at a time, we tested each SNP separately and applied Bonferroni adjustment. We applied and compared with the adapted versions of Huang *et al.*, CaMMEL and LASSO because the corresponding original methods only test θ in the outcome model. For all the adapted versions, we utilize SKAT to test β in the mediator model to be maximally comparable with our SMUT results. In other words, adapted Huang *et al.* is SKAT + original Huang *et al.* with SKAT corresponding to the testing strategy in the mediator model and original Huang *et al.* to the testing strategy in the outcome model. Similarly, for CaMMEL and LASSO, we use adapted CaMMEL and SKAT+CaMMEL exchangeably; adapted LASSO and SKAT+CaMMEL exchangeably.

As detailed above, we simulated causal SNPs only from the pool of common (MAF $\geq$ 1%) SNPs. By default, we tested all common SNPs in the region to mimic the realistic scenario where we have relatively little information regarding which SNPs are causal, at an established GWAS locus. To test the robustness and generalizability of the methods, we considered two alternative testing strategies each with a reduced set of genetic variants modeled. For the first testing strategy, we assume prior knowledge of eQTL SNPs (union of shared and mediator specific causal SNPs) and test only these eQTL SNPs. On the positive side, such an approach results in a reduced model with causal SNPs considered only. On the negative side, a subset of causal markers (specifically, the outcome specific causal SNPs) are not modeled. The second strategy tests SNPs with MAF $\geq$ 5%, thus missing true causal SNPs with MAF between 1 and 5%.

3 Results

3.1 Type-I error in simulations

We evaluated SMUT along with alternative methods in simulations. SMUT manifested controlled Type-I error rates, at $\alpha = 0.05$ level, regardless of causal SNP density, as shown in Figures 2 and 3 for sparse and dense scenarios, respectively. Note that the first panel ($c_\beta = 0$) and the left-most point ($\theta = 0$) in other panels ($c_\beta \neq 0$) all correspond to the null of no mediation through the mediator. Adapted Huang *et al.*'s method also showed protected Type-I error. In contrast, Sobel test and SMR showed substantial inflation in Type-I error, particularly when c_β is large. For example, when $c_\beta = 0.2$, $\theta = 0$ and sparse causal SNPs, Type-I error rates for Sobel test and SMR are 90% and 100%, respectively. Such marked inflation in Type-I error is likely due to the more severe violation of the assumption of no pleiotropy, made by these two methods, as c_β increases. Adapted CaMMEL also showed Type-I error inflation. For example, among sparse causal SNPs when $c_\beta = 0.2$, $\theta = 0$, the Type-I error rate is 100%. We suspect such inflation is due to the

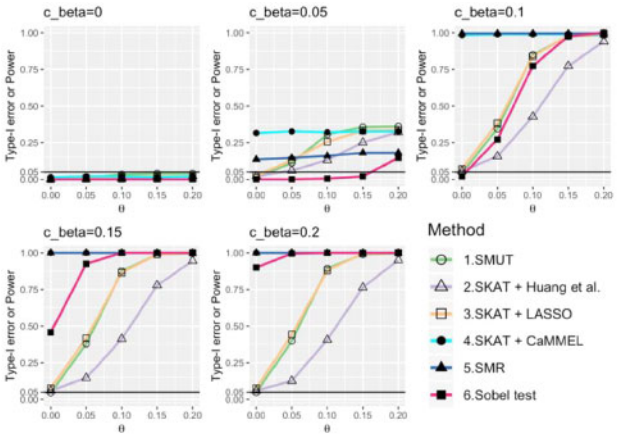

Fig. 2. Power and Type-I error under sparse causal SNPs scenario. The *x*-axis is the true mediator effect(θ) on the outcome. The *y*-axis is the power or Type-I error. Sub-figures vary in c_β value. $c_\beta = 0$ (top-left sub-figure) or $\theta = 0$ (left-most points in each sub-figure) are null settings where *y*-axis represents the corresponding Type-I error. When $c_\beta \neq 0$ and $\theta \neq 0$, it is under alternative hypothesis and *y*-axis represents the corresponding power

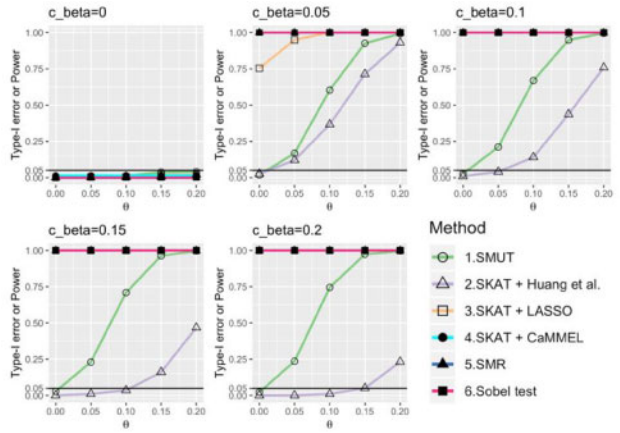

Fig. 3. Power and Type-I error under dense causal SNPs scenario. *X*-axis and *y*-axis are the same as in Figure 2

fact that CaMMEL was developed for joint testing of multiple mediators via a Bayesian framework to borrow information across mediators. Thus, when testing one single mediator, lack of information in the Bayesian inference can lead to Type-I error inflation. Adapted LASSO had severe Type-I error inflation when the causal SNPs were dense (Fig. 3). For instance, when $c_\beta = 0.05$ and $\theta = 0$ Type-I error rate is 75%. This is likely due to the violation of LASSO's sparsity assumption (Zhao and Yu, 2006).

Assuming normality of $\gamma_j (j = 1, 2, \ldots, q)$ in the outcome model may not be strictly correct when some SNPs are non-causal (γ_j exactly zero) while others are causal. A mixture distribution would be more appropriate. But our approach gives valid tests in simulations even when the assumption may not be valid (Supplementary Figs S1 and S2).

3.2 Power in simulations

We assessed power only for tests with protected Type-I error, namely our SMUT and adapted Huang *et al.* SMUT demonstrated large power gains when the causal SNPs were either sparse or dense. For example, dense causal SNPs when $c_\beta = 0.2$, $\theta = 0.15$, SMUT and

adapted Huang *et al.* had 97% and 5% power, respectively and the power gain was 92%. Power gains appeared more profound with increasing c_β, likely because adapted Huang *et al.* became very conservative when the pleiotropy effect (c_β) was large. Details and results of testing robustness of SMUT are in Supplementary Section 3.

3.3 Real data application: METSIM dataset

The METSIM study is a population-based study with 10 197 males, aged 45–73 years, randomly selected from the population register of Kuopio town in eastern Finland (population 95 000) (Stancakova *et al.*, 2009). We analyzed genotype, gene expression and phenotype data in the subset of 770 participants with gene expression measurements from subcutaneous adipose tissue (Civelek *et al.*, 2017). The outcome of interest is plasma adiponectin levels. All METSIM subjects participated in a 1-day outpatient visit to the Clinical Research Unit at the University of Kuopio for data collection, which included an interview for their medical history and a blood sample following a 12-h fast. Plasma was measured using the Human Adiponectin Elisa kit (LINCO Research).

Here, we tested two 'positive control' loci for which our previous study (Civelek *et al.*, 2017) provided mechanistic evidences. The first locus was the adiponectin-associated GWAS locus *ARL15* (with the index SNP rs6450176 being an *ARL15* intronic variant), where the association might be mediated, at least in part, through altered expression of the *FST* gene located further (>521 kb from rs6450176) away instead of *ARL15* (Civelek *et al.*, 2017; Martin *et al.*, 2017). The second locus was the *ADIPOQ* locus, also associated with adiponectin levels.

We first extracted SNPs within ±1 Mb of the corresponding genes, *ARL15* union *FST* and *ADIPOQ* union *ADIPOQ-AS1* for the two loci, respectively. In terms of phenotypic outcome, namely adiponectin, trait levels were inverse normal transformed after adjusting for age and BMI, following our previous work (Civelek *et al.*, 2017). For the first *ADIPOQ* locus, we tested 286 SNPs with adiponectin association *P*-value $<5\times10^{-8}$, using SMUT, adapted Huang *et al.*'s method, adapted CaMMEL, CIT, SMR and Sobel test. Results are summarized in Table 2. Huang *et al.*'s method returned no results (therefore not shown in Table 2) because it required standardized genotype data which can be undefined for low frequency SNPs. SMUT and SMR both showed significant mediation effects through *ADIPOQ* on adiponectin: SMUT for two probesets and SMR for two probesets. For the second *FST-ARL15* locus, we tested 366 SNPs with MAF ≥ 1% and adiponectin association *P*-values <0.01. Only SMUT detected significant mediation effects through *FST* (but not *ARL15*) on the adiponectin. These results suggest that our SMUT is more powerful for detecting genuine mediation effects.

4 Discussion

We propose SMUT, a flexible regression-based approach that tests the joint mediation effects of multiple genetic variants on an outcome through a given mediator (e.g. gene). We demonstrate, through extensive simulations, that SMUT preserves Type-I error rate. Our IUT approach essentially takes the maximum of the *P*-values from separately testing β being zero and θ being zero, with the Type-I error for the likely more influential β part protected by the well-established SKAT method. More stringent filtering can be applied by adopting multiple testing adjustments such as Bonferroni or FDR correction. SMUT is statistically more powerful than alternative methods including adapted Huang *et al.*'s method, adapted LASSO, adapted CaMMEL, Sobel test and SMR.

SMUT has several major advantages over alternative methods. First, as a regression-based approach under the mediation analysis framework, SMUT can distinguish mediation from pleiotropy. Second, SMUT generalizes the framework of Baron and Kenny to multiple genetic variants, while methods including SMR and Sobel test can only test one single variant at a time. Third, SMUT naturally accommodates correlation (or LD) among genetic variants while many methods including MR-Egger assume genetic variants under testing are uncorrelated. Fourth, SMUT enables relatively large number of SNPs by fitting a mixed effects model, while sparse fixed effects model (e.g. LASSO) relies on sparsity of the true causal SNPs and may cause inflated Type-I error if violating the sparsity assumption. Finally, SMUT, even its present form, can handle mediators other than gene expression (as presented in the manuscript). For example, molecular measurements such as chromatin spatial organization, histone modification, transcription factor binding affinity and protein abundance can all serve as valid mediators (Schmitt *et al.*, 2016; Sun *et al.*, 2016; The GTEx Consortium, 2015; Xu *et al.*, 2016).

Conceptually, TWAS methods are also designed to elucidate mechanisms regarding the mediation effects of multiple SNPs via gene expression on phenotypic outcome. However, as previously mentioned, TWAS is designed for scenarios where eQTL and GWAS

Table 2. Results from the METSIM study

			P-values				
Probesets	No. of SNPs	Gene	SMUT	Adapted CaMMEL	CIT	SMR	Sobel test
11734558_a_at	286	*ADIPOQ*	0.07	0.09	1.0	0.08	0.07
11734559_x_at	286	*ADIPOQ*	**0.01**	0.09	0.54	**0.03**	0.07
11734560_x_at	286	*ADIPOQ*	0.90	0.09	1.0	0.08	1.0
11752564_x_at	286	*ADIPOQ*	**0.04**	0.09	0.89	**0.03**	0.27
11724032_a_at	366	*FST*	**0.02**	0.09	1.0	1.0	1.0
11732712_a_at	366	*FST*	**0.01**	0.09	1.0	1.0	1.0
11732713_at	366	*FST*	**0.03**	0.09	1.0	1.0	1.0
11731654_at	366	*ARL15*	1.0	0.09	1.0	1.0	1.0
11757014_a_at	366	*ARL15*	0.13	0.09	1.0	1.0	1.0

Note. We used SMUT and other alternative methods (adapted CaMMEL, CIT, SMR and Sobel test) to test two loci, the *ARL15* locus and the *ADIPOQ* locus. SNPs within corresponding genes, *ARL15* union *FST* and *ADIPOQ* union *ADIPOQ-AS1* for the two loci respectively, are extracted. For the *ADIPOQ* locus, both SMUT and SMR showed significant mediation effects through *ADIPOQ* on adiponectin. For the *ARL15* locus, only SMUT detected significant mediation effects through *FST* (but not *ARL15*) on the adiponectin. The *P*-values are adjusted using Bonferroni correction. Numbers in bold are the P-values less than 0.05.

datasets are from two separate sets of study participants. Our SMUT method is designed for the scenario where we have genotype, mediator and phenotype information measured in the same study subjects. Therefore, we have not directly compared with TWAS methods and deem our SMUT and TWAS useful for different data scenarios.

SMUT can be further extended in several directions. It can be extended to accommodate binary, survival or longitudinal phenotypic outcome, given its regression-based framework. These extensions, however, are non-trivial because the outcome model will be a generalized linear mixed model with random effects (for SNPs) that are high dimensional, and are shared across samples unlike in standard repeated measures settings. These complexities entail the exploration of Laplace approximation of the likelihood or partial likelihood function for proper and computationally tractable testing of theta, which we are actively pursuing and warrants separate publication. We can also extend SMUT to simultaneously model multiple mediators, which may yield improved power for testing at the price of stronger modeling assumptions.

With more genotyping-based GWAS and large whole genome sequencing efforts underway, the already dauntingly large number of GWAS variants will continue to increase. Approaches generating hypotheses on the mechanisms underlying these variants are imperative. We anticipate SMUT will be a powerful tool in this post-GWAS era to help with bridging the functional gap of GWAS, prioritizing functional follow-up and disentangling the potential causal mechanism from DNA to phenotype for a new drug discovery and personalized medicine.

Acknowledgements

We thank the METSIM individuals and investigators to share the data. We also thank the reviewers for helpful comments and suggestions.

Funding

This work was supported by the National Institutes of Health [R01HL129132 to Y.L., R01DK093757 to K.L.M.].

Conflict of Interest: none declared.

References

Ainsworth,H.F. *et al.* (2017) A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements. *Genet. Epidemiol.*, **41**, 577–586.

Barfield,R. *et al.* (2017) Testing for the indirect effect under the null for genome-wide mediation analyses. *Genet. Epidemiol.*, **41**, 824–833.

Barfield,R. *et al.* (2018) Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genet. Epidemiol.*, **42**, 418–433.

Baron,R.M. and Kenny,D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, **51**, 1173–1182.

Berger,R.L. and Hsu,J.C. (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.*, **11**, 283–319.

Bowden,J. *et al.* (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.*, **44**, 512–525.

Civelek,M. *et al.* (2017) Genetic regulation of adipose gene expression and cardio-metabolic traits. *Am. J. Hum. Genet.*, **100**, 428–443.

Civelek,M. and Lusis,A.J. (2014) Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.*, **15**, 34.

Engle,R.F. (1984) Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handb. Econom.*, **2**, 775–826.

Gamazon,E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.

Gusev,A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245.

Huang,Y.-T. *et al.* (2014) Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat.*, **8**, 352–376.

Huang,Y.T. (2015) Integrative modeling of multi-platform genomic data under the framework of mediation analysis. *Stat. Med.*, **34**, 162–178.

Ionita-Laza,I. *et al.* (2013) Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.*, **92**, 841–853.

Lawlor,D.A. *et al.* (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.*, **27**, 1133–1163.

Lee,S. *et al.* (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.

Lin,X. (1997) Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309–326.

Lloyd-Jones,L.R. *et al.* (2017) The genetic architecture of gene expression in peripheral blood. *Am. J. Hum. Genet.*, **100**, 228–237.

MacKinnon,D.P. *et al.* (2007) Mediation analysis. *Annu. Rev. Psychol.*, **58**, 593–614.

Martin,J.S. *et al.* (2017) HUGIn: Hi-C unifying genomic interrogator. *Bioinformatics*, **33**, 3793–3795.

Millstein,J. *et al.* (2009) Disentangling molecular relationships with a causal inference test. *BMC Genet.*, **10**, 23.

Radhakrishna Rao,C. and Bartlett,M.S. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Math. Proc. Cambridge Philos. Soc.*, **44**, 50.

Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.

Schmitt,A.D. *et al.* (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.*, **17**, 2042–2059.

Smith,G.D. and Ebrahim,S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.*, **32**, 1–22.

Sobel,M.E. (1982) Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.*, **13**, 290–312.

Solovieff,N. *et al.* (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, 483–495.

Stancakova,A. *et al.* (2009) Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6, 414 Finnish men. *Diabetes*, **58**, 1212–1221.

Sun,W. *et al.* (2016) Common genetic polymorphisms influence blood biomarker measurements in COPD. *PLoS Genet.*, **12**, 1–33.

The GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B*, **58**, 267–288.

VanderWeele,T.J. (2016) Mediation analysis: a practitioner's guide. *Annu. Rev. Public Health*, **37**, 17–32.

Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Xu,Z. *et al.* (2016) A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics*, **32**, 650–656.

Yang,F. *et al.* (2017) Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res.*, **27**, 1859–1871.

Zhao,P. and Yu,B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zhu,Z. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.