BMC Bioinformatics

RESEARCH

Open Access

Estimating the effect size of a hidden causal factor between SNPs and a continuous trait: a mediation model approach

Zhuoran Ding[1], Marylyn D. Ritchie[1,2,3], Benjamin F. Voight[2,3,4,5*†] and Wei-Ting Hwang[1*†]

†Benjamin F. Voight and Wei-Ting Hwang jointly supervised this work

*Correspondence: bvoight@pennmedicine.upenn.edu; whwang@pennmedicine.upenn.edu

[1] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA
[2] Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA
[3] Institude for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA
[4] Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA
[5] Institute of Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Background: Observational studies and Mendelian randomization experiments have been used to identify many causal factors for complex traits in humans. Given a set of causal factors, it is important to understand the extent to which these causal factors explain some, all, or none of the genetic heritability, as measured by single-nucleotide polymorphisms (SNPs) that are associated with the trait. Using the mediation model framework with SNPs as the exposure, a trait of interest as the outcome, and the known causal factors as the mediators, we hypothesize that any unexplained association between the SNPs and the outcome trait is mediated by an additional unobserved, hidden causal factor.

Results: We propose a method to infer the effect size of this hidden mediating causal factor on the outcome trait by utilizing the estimated associations between a continuous outcome trait, the known causal factors, and the SNPs. The proposed method consists of three steps and, in the end, implements Markov chain Monte Carlo to obtain a posterior distribution for the effect size of the hidden mediator. We evaluate our proposed method via extensive simulations and show that when model assumptions hold, our method estimates the effect size of the hidden mediator well and controls type I error rate if the hidden mediator does not exist. In addition, we apply the method to the UK Biobank data and estimate parameters for a potential hidden mediator for waist-hip ratio beyond body mass index (BMI), and find that the hidden mediator has a large effect size relatively to the effect size of the known mediator BMI.

Conclusions: We develop a framework to infer the effect of potential, hidden mediators influencing complex traits. This framework can be used to place boundaries on unexplained risk factors contributing to complex traits.

Keywords: Mediation, Gaussian mixture model, Markov-Chain Monte-Carlo, Obesity

Introduction

One of the goals in studying the associations between heritable traits and disease outcomes is to identify which of these factors are truly causal for the outcome. Expansion of causal inference studies like Mendelian randomization studies [1] in recent years have provided one piece of causal evidence to risk factors identified from observational

epidemiological studies [2]. However, generating causal evidence between exposures and disease outcome leaves unaddressed the question of whether the genetic heritability of the trait can be fully explained by the set of known causal factors or if there exist additional ones that are unidentified but further explain disease risk or trait heritability. For example, it is known that high body mass index (BMI) and potentially low-density lipoprotein cholesterol (LDL-C) increase the risk of developing type 2 diabetes (T2D) [3, 4], but can the genetic heritability of T2D be fully explained by BMI and LDL-C? If the heritability of a disease can be fully explained by the set of existing known causal factors, then research on the disease can focus on studying the biological mechanisms of those causal factors. On the other hand, if there is genetic heritability that remains unexplained, other causal factors for the disease may exist and are currently hidden from observation. Studying the characteristics of the hidden causal factors may provide insights into the novel biological associations and mechanisms that remain undiscovered.

The classic mediation framework decomposes the associations between the exposures and the outcome into direct effect and indirect effect through a mediator (i.e., mediated effects) [5, 6]. The question described earlier can be considered within a mediation framework with the trait of interest as the outcome, identified single nucleotide polymorphisms (SNPs) associated with the trait as the exposure, and the known causal factors of the trait as the mediators. Under this framework, any remaining direct effects between the identified SNPs and the outcome trait are viewed as the residual associations that are not explained by causal factors or mediators included in the model. Thus, the residual associations could be due to one or more hidden mediators. As the first step towards learning about the residual associations in the following work, we consider the simplest case that the remaining direct effects between the identified SNPs and the outcome trait are due to a single hidden mediator. Under this case, we can further decompose the unexplained genetic heritability into two parts: (i) the SNP effects on the hidden mediator, and (ii) the effect size of the hidden mediator on the outcome trait. In this study, we aim to determine whether the hidden mediator exists, and if so, estimate the effect size of the hidden mediator on the outcome trait. We believe that the results of our work can be helpful in at least two ways. First, the compatibility of the data with the presence of a hidden mediator suggests additional work to understand complex trait heritability. Secondly, given the presence of a hidden mediator, enumerating the distribution of effects of that mediator across loci could be important. For example, in T2D, a locus which is entirely mediated by BMI / obesity (e.g., FTO) but not by other risk factors is interesting, and could point to known disease etiology [7]. In contrast, a locus which is not explained at all (or incompletely) by known mediators could help focus investigation on a locus where novel biological insight could be obtained.

In this work, we assume that the values of the SNPs, the known mediators, and the outcome trait are accurately measured and there are no unmeasured confounding variables in the model. Furthermore, we assume that the effect sizes of the SNPs on the standardized known and hidden mediators have the same distribution. This is reasonable due to the following: because there is little or no prior information about the hidden mediator, the unexplained genetic heritability can theoretically be decomposed by infinitely many combinations of the two parts. Thus, to limit the number of possible combinations of the two parts and infer a reasonable range of the hidden mediator's effect

size on the outcome, we propose to restrict the effect sizes of the SNPs on the hidden mediator to be similar to the effect sizes of the SNPs on the known mediators. Under this assumption, one part of the unexplained genetic heritability, that is, the SNP effects on the hidden mediator, can be learned from the SNP effects on the known mediators. Furthermore, the effect size of the hidden mediator on the outcome trait can be inferred by dividing the direct effects between the SNPs and the outcome trait by the inferred SNP effects on the hidden mediator.

The rest of the report is organized as follows: In the "Methods" Section, we provide a broad overview of our approach and describe each step of our method in detail. We also describe the settings for the simulation study. In the "Results" Section, we present the simulation results and the application on investigating the trait of waist-hip ratio. We conclude inthe "Discussion" Section with thoughts about limitations and possible extensions to the approach.

Methods

The mediation model and notations

We illustrate our method using a model with two known mediators, although our method can extend to the case with more than two mediators or only one known mediator. We denote a continuous trait of interest as Y, the vector of SNPs associated with Y as $\boldsymbol{G}$, and the two known mediating causal factors of Y as M_1 and M_2 (Fig. 1A). M_1 and M_2 are both standardized to have unit variance. Furthermore, the SNP effects from G to M_1 are represented by a vector $\boldsymbol{a_1}$ and the SNP effects from $\boldsymbol{G}$ to M_2 are represented by a vector $\boldsymbol{a_2}$, the direct effects between $\boldsymbol{G}$ and Y are represented by a vector $\boldsymbol{c}$, the effect size of M_1 on Y is denoted by a scalar b_1, and the effect size of M_2 on Y is denoted by a scalar b_2.

If there is an unexplained genetic heritability between $\boldsymbol{G}$ and Y (i.e., $c \neq 0$), then we assume a hidden mediator exists and is denoted by M_H (Fig. 1B). The SNP effects from $\boldsymbol{G}$ to M_H are represented by a vector $\boldsymbol{a_H}$. The goal of the proposed work is to infer the effect size of the hidden mediator M_H on Y, denoted by b_H. To account for the scenario that some of the SNPs in $\boldsymbol{G}$ are not associated with M_H, we use π_H to denote the proportion of the SNPs in $\boldsymbol{G}$ that are associated with M_H.

If we were to observe M_H, the direct effect vector c can be decomposed as $\boldsymbol{c} = \boldsymbol{a_H} b_H$. However, since M_H is not observed, we can only estimate $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, b_1, b_2, and $\boldsymbol{c}$ as shown in

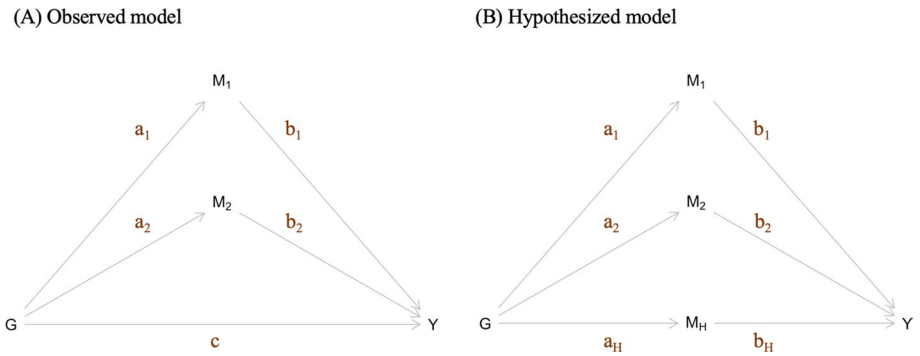

Fig. 1 The mediation model framework. (**A**) The observed model. (**B**) The hypothesized model

Fig. 1A but not $\boldsymbol{a_H}$ and b_H as shown in Fig. 1B. We propose to use the estimates of $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, and $\boldsymbol{c}$, denoted as $\boldsymbol{a^*_1}$, $\boldsymbol{a^*_2}$, and $\boldsymbol{c^*}$ to infer $\boldsymbol{a_H}$ and, subsequently, infer b_H. For simplicity, we denote the joint vector of $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ and the joint vector of $\boldsymbol{a^*_1}$ and $\boldsymbol{a^*_2}$ as $\boldsymbol{a}$ and $\boldsymbol{a^*}$, respectively.

Overview of the proposed method and the rationales

To infer the effect size of the hidden mediator, b_H, we utilize the fact that we can decompose the direct effect $\boldsymbol{c}$ as $\boldsymbol{a_H} b_H$. To do so, we assume that the SNP effects on the hidden mediator, $\boldsymbol{a_H}$, share some similarities with the SNP effects on the known mediators, $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$. We first consider the simplest case, in which we assume that all the SNP effects (on both the known and unknown mediators) come from the same distribution. Under this assumption, the true mean and standard deviation of $\boldsymbol{a^*_H}$ can be consistently estimated by the SNP effects $\boldsymbol{a^*_1}$ and $\boldsymbol{a^*_2}$ if the sample size is large and a large number of SNPs are included in the model. We estimate the SNP effects $\boldsymbol{a^*_1}$ and $\boldsymbol{a^*_2}$ by fitting two linear regression models with the known mediators M_1 and M_2 as the dependent variables and $\boldsymbol{G}$ as the independent variable. Recognizing that assuming similar SNP effect sizes and variances on different mediators may be a strong assumption, we present a more general setting in which the SNP effects could vary according to a three-level structure (Additional file 1).

One challenge in the decomposition of the direct effects of SNPs on Y is that we might not expect every SNP associated with the outcome trait will be associated with the known and hidden mediators. Therefore, we model the SNP effects on the known mediators ($\boldsymbol{a}$), using a mixture model with a point mass at zero and a true effect size distribution that centers at a non-zero value. Because we estimate the SNPs effects on the known mediators ($\boldsymbol{a^*}$) using linear regression models that come with estimation uncertainty, the distribution of $\boldsymbol{a^*}$ will be a mixture of a distribution centered at zero and the true effect distribution with a non-zero mean and a variance that is larger than the true dispersion. Similarly, the SNP effects on the hidden mediator, $\boldsymbol{a_H}$, can also be modeled using a mixture model with a point mass at zero and a true effect distribution that centers at a non-zero value. For the same reason as in the case of $\boldsymbol{a^*}$, the estimated $\boldsymbol{c^*}$ will be the mixture of a point mass at zero and a true effect distribution not centered at zero and with a larger dispersion. Therefore, we utilize Gaussian mixture models (GMMs) to model the distributions of $\boldsymbol{a^*}$ and $\boldsymbol{c^*}$.

The proposed multi-step method

Our method consists of three major steps as shown in Fig. 2. In Step 1, we estimate the individual SNP effects on the known mediators to obtain $\boldsymbol{a^*}$ and estimate the direct effects between the SNPs and the outcome trait to obtain $\boldsymbol{c^*}$ by fitting a series of linear regression models. In Step 2, we fit GMMs on $\boldsymbol{a^*}$ and $\boldsymbol{c^*}$ using an Expectation–Maximization (EM) algorithm to separate the SNP effects on the known and hidden mediators from the zero-mean noises and estimate the GMM parameters from both distributions. In Step 3, we incorporate the estimated GMM parameters from Step 2 to a GMM Markov Chain Monte Carlo (MCMC) procedure to generate a posterior distribution for

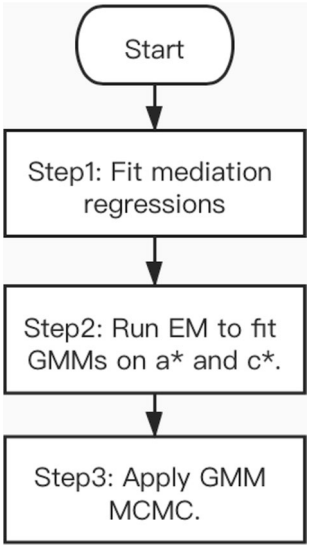

Fig. 2 Flowchart of the multi-step method. In the first step, we use linear regression models to estimate the SNP effects on the known mediators ($\boldsymbol{a}^*$) and the direct effects between the SNPs and the outcome ($\boldsymbol{c}^*$). In the second step, we apply the EM algorithm to fit GMMs on $\boldsymbol{a}^*$ and $\boldsymbol{c}^*$. In the third step, a MCMC procedure is performed using the estimated GMM parameters from the last step in the priors to generate a posterior distribution for the hidden mediator's effect size, b_H

b_H. Details of each step are presented below. All steps are implemented in *R* (version 3.6.1) [8].

Step 1: Mediation regressions

In Step 1, we estimate the SNP effects on the known mediators, $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$, and direct effects between the SNPs by fitting linear regressions for each known mediator separately where the mediator (e.g., M_1) is the dependent variable and the elements of $\boldsymbol{G}$ are the independent variables. To estimate the direct effects between the SNPs and the outcome trait, $\boldsymbol{c}$, we fit a linear regression with Y being the dependent variable and $\boldsymbol{G}$, M_1, M_2, and other covariates being the independent variables. The resulting estimated effects are $\boldsymbol{a}^*$ and $\boldsymbol{c}^*$. We denote $\boldsymbol{a}^*$ as the joint vector of $\boldsymbol{a}^*_{\boldsymbol{1}}$ and $\boldsymbol{a}^*_{\boldsymbol{2}}$.

Step 2: EM

In Step 2, we separate the true effects in $\boldsymbol{c}^*$ from the zero-mean noise component by fitting a GMM. In addition, because it is possible that not all the SNPs are associated with the set of known mediators, we also fit a GMM on $\boldsymbol{a}^*$ to capture the actual effects of the SNPs on the known mediators. Specifically, we use the EM algorithm to fit the GMMs on $\boldsymbol{a}^*$ and $\boldsymbol{c}^*$ via the *normalmixEM* function in the *R* package *mixtools* (version 1.2.0) [9–11] with the initial value of mixing proportions, lambda, set to 0.5, which represents that initially, the SNPs have equal probabilities of being associated with the trait or not (i.e., the hidden mediator in the case of $\boldsymbol{c}^*$ or the corresponding known mediators in the case of $\boldsymbol{a}^*$).

The EM algorithm works as follows. Let X_i for $i = 1 \ldots n$ be random variables generated from a GMM consist of two normal distributions (Eq. 1), and let $Z_i = \{1,2\}$ for $i = 1 \ldots n$

be binary latent variables that each indicates which of the two normal distribution the corresponding X_i comes from.

Let $\theta^{(r)} = \{\pi_1^{(r)}, \mu_1^{(r)}, \sigma_1^{2,(r)}, \pi_2^{(r)}, \mu_2^{(r)}, \sigma_2^{2,(r)}\}$ be the set of GMM parameters computed at iteration r.

At iteration r, during the E step, compute $E\left(Z_i(k)|x_i, \theta^{(r-1)}\right) = P\left(Z_i = k|x_i, \theta^{(r-1)}\right)$, where $k = \{1,2\}$ as in Eq. 2. During the M step, compute elements of $\theta^{(r)}$ as in Eqs. 3–5.

$$P(X = x) = \pi_1 N\left(x|\mu_1, \sigma_1^2\right) + \pi_2 N\left(x|\mu_2, \sigma_2^2\right), \textit{ where } \pi_1 + \pi_2 = 1 \tag{1}$$

$$P(Z_i = k|x_i) = \frac{\pi_k^{(r-1)} N\left(x_i|\mu_k^{(r-1)}, \sigma_k^{2,(r-1)}\right)}{\pi_1 N\left(x_i|\mu_1^{(r-1)}, \sigma_1^{2,(r-1)}\right) + \pi_2 N\left(x_i|\mu_2^{(r-1)}, \sigma_2^{2,(r-1)}\right)} \tag{2}$$

$$\pi_k^{(r)} = \frac{\sum_{i=1}^n E\left(Z_i(k)|x_i, \theta^{(r-1)}\right)}{n} \tag{3}$$

$$\mu_k^{(r)} = \frac{\sum_{i=1}^n E\left(Z_i(k)|x_i, \theta^{(r-1)}\right) x_i}{\sum_{i=1}^n E\left(Z_i(k)|x_i, \theta^{(r-1)}\right)} \tag{4}$$

$$\sigma_k^{2,(r)} = \frac{\sum_{i=1}^n E\left(Z_i(k)|x_i, \theta^{(r-1)}\right)\left(x_i - \mu_k^{(r)}\right)^2}{\sum_{i=1}^n E\left(Z_i(k)|x_i, \theta^{(r-1)}\right)} \tag{5}$$

The fitted GMM should consist of a near zero-mean normal distribution, which are due to the zero-mean noise and a non-zero-mean normal distribution, which are due to the true effects. The means and the standard deviations of $\boldsymbol{a}^*$'s and $\boldsymbol{c}^*$'s true effect component can be estimated by the means and the standard deviations of the non-zero-mean distributions from the fitted GMMs. To avoid occasional convergence issues and extreme estimates, we run the EM algorithm multiple times and take the median value of the estimated true effect distribution means from all the runs. Based on our experience, generating 15 runs of the EM GMM is typically sufficient to obtain a well-fitted model.

Step 3: GMM

In Step 3, we specify a GMM as shown in Eq. 6 and perform a MCMC procedure to generate a posterior distribution for b_H using the R package *rjags* (version 4.10) [12, 13].

$$c_j \sim \begin{cases} N\left(0, \sigma_1^2\right) \textit{ if } z_j = 1 \\ N\left(b_H \mu_a, \sigma_2^2\right) \textit{ if } z_j = 2 \end{cases} \tag{6}$$

The data input into the GMM MCMC procedure are the elements of $\boldsymbol{c}^*$. We denote the elements of $\boldsymbol{c}^*$ as c_j for $j = 1 \ldots n$, where n is the number of SNPs in the model. μ_a, σ_1^2, and σ_2^2 are constants in the model. The value of μ_a is set to the estimated mean of the true effect Gaussian distribution of the fitted GMM on $\boldsymbol{a}^*$ in Step 2; the value of σ_1^2 is set to the estimated variance of the zero-mean Gaussian distribution in the fitted GMM in step

2; the value of σ_2^2 is set to the estimated variance of the non-zero-mean Gaussian distribution in the fitted GMM in step 2. The parameters estimated by this MCMC procedure are b_H and the binary indicator variables z_j for $j=1...n$, which indicate which Gaussian distribution in the GMM the corresponding c_j belongs to. We specify the same categorical distribution prior for z_j, where the weight parameters of the two categories are set to the corresponding estimated weights for the two Gaussian distributions in the fitted GMM in Step 2. Lastly, we specify a normal prior with mean zero and variance 100 for b_H so its distribution is almost flat at small values close to 0, which is the potential region of the hidden mediator's effect size.

When b_H is equal to zero, the distribution for $\boldsymbol{c}^*$ reduces to a single Gaussian distribution instead of a GMM, and the resulting GMM from EM will likely assign a very small weight to one of the Gaussian distributions in the GMM. If the Gaussian distribution that involves b_H in the MCMC procedure happens to be the one that receives a very small weight, the posterior distribution of b_H will span a very wide region around the true value of b_H. This is because during each MCMC iteration, due to the small weight, very few or none of the elements in $\boldsymbol{c}^*$ will be assigned to the Gaussian distribution that involves b_H so that the MCMC procedure is uncertain about the estimation of b_H. This situation can also occur when the true value of b_H is very small such that the two Gaussian distributions in the GMM are not separable. Under these scenarios, the resulting interval estimators will be extremely wide and will not be useful in terms of giving a precise estimate of b_H. Thus, if one observes an extremely wide posterior distribution of b_H, we propose to flip the binary labeling of the fitted GMM from Step 2 and perform Step 3 again. Based on the results of the simulation study presented below, we observe that this adjustment generally results in more meaningful estimates of b_H.

A simulation study

We conduct a simulation study to evaluate the proposed method. We consider a base case and eight additional settings (Table 1). The base case is an ideal setting for our method. For each of the eight settings, we vary different aspects of the base case and evaluate the behavior of our method. In Setting 1, we vary the proportion of the SNPs that are associated with the hidden mediator, which we denote as π_H; in Setting 2, we allow two of the known mediators to have negative effects on the outcome trait; in Setting 3, we decrease the sample size (number of individuals); in Setting 4, we consider cases where there are 1 and 10 known mediators of the outcome trait; in Setting 5, we consider four cases where some of the known mediators affect other known mediators; in Setting 6, we simulate two cases under the three-level SNP effect structure; in Setting 7, we simulate the case where b_H is equal to zero (i.e., negative control); in Setting 8, we vary the number of SNPs included in the model. For the base case and Settings 1–7, we consider a scenario where there are 70 SNPs associated with the outcome trait and another scenario where there are 500 SNPs associated with the outcome trait. We select the 500 SNP scenario because it is similar to the number of SNPs used in the data application example (522 SNPs) described in the next section. We also examine the 70 SNP scenario and repeated all the simulations done with 500 SNPs to access how our method behaves when there are a significantly smaller number of SNPs in the model.

Table 1 Parameters specifications of the 9 simulation settings

Setting	π_H	Know mediator effects	Sample size	Number of known mediators	Associations among mediators	Level 1 and Level 2 standard deviation ratios	Hidden mediator effect	Number of SNPs
Base case	0.8	(0.4, 0.2, 0.3, 0.2, 0.4)	100,000	5	Independent	0:1	Non-zero	70; 500
Setting 1	0.3; 0.5; 1	(0.4, 0.2, 0.3, 0.2, 0.4)	100,000	5	Independent	0:1	Non-zero	70; 500
Setting 2	0.8	(0.4, 0.2, 0.3, -0.2, -0.4)	100,000	5	Independent	0:1	Non-zero	70; 500
Setting 3	0.8	(0.4, 0.2, 0.3, 0.2, 0.4)	25,000; 50,000	5	Independent	0:1	Non-zero	70; 500
Setting 4	0.8	(0.4); (0.4, 0.2, 0.3, 0.2, 0.4, 0.2, 0.1, 0.3, 0.2, 0.2)	100,000	1; 10	Independent	0:1	Non-zero	70; 500
Setting 5	0.8	(0.4, 0.2, 0.3, 0.2, 0.4)	100,000	5	M_1->(0.3) M_2; M_1->(0.9) M_2; M_1->(0.3) M_2 and M_3->(0.2) M_4; M_1->(0.5) M_2 and M_3->(0.4) M_4	0:1	Non-zero	70; 500
Setting 6	0.8	(0.4, 0.2, 0.3, 0.2, 0.4)	100,000	5	Independent	1:3; 1:1	Non-zero	70; 500
Setting 7	/	(0.4, 0.2, 0.3, 0.2, 0.4)	100,000	5	Independent	0:1	Zero	70; 500
Setting 8	0.8	(0.4, 0.2, 0.3, 0.2, 0.4)	100,000	5	Independent	0:1	Non-zero	20; 40; 700

Parameters separated by ";" belong to separate simulations

For both scenarios, we let the hidden mediator's effect size on the outcome trait, b_H, be 0.02, 0.25, and 0.5 and perform 1000 independent simulations for each of the three values of b_H. Also, for Settings 1–6, we simulate data for 49 additional b_H between 0.02 and 0.5 with a step size of 0.01 to show a behavior trend of our posterior distribution as b_H increases. As previously mentioned, when the true value of b_H is very small or equal to zero, the MCMC posterior distribution of b_H may be too wide occasionally to make any meaningful inference about its true value. Therefore, we apply the labeling-switching adjustment if the width of the 90% quantile interval derived from the initial posterior distribution for b_H is wider than 5. The detailed setting of the base case is presented in the next paragraph with its simulation results being presented in "Results" Section. The detailed settings and the simulation results of the other eight settings are presented in the Additional file 1. All simulations were performed in *R* (version 3.6.1) [8].

The data for the base case are simulated as follows. The sample size is 100,000. The SNPs associated with the outcome trait, $\boldsymbol{G}$, are simulated independently with minor allele frequencies generated from a uniform distribution between 0.1 and 0.5, and we assume the SNP effects are additive. The SNP effects on the mediators $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, $\boldsymbol{a_3}$, $\boldsymbol{a_4}$, $\boldsymbol{a_5}$ and $\boldsymbol{a_H}$ are generated from mixture distributions of zero point-masses and normal distributions with the mean being 0.2 and the standard deviation being 0.08 as shown in Eqs. 7–12. The frequencies that the five known mediators are associated with the exposure SNPs are (0.5, 0.6, 0.8, 0.2, 0.5), i.e., 50% of the SNPs associated with the outcome trait are associated with M_1; 60% of the SNPs associated with the outcome trait are associated with M_2, and so forth. There are five known mediators of the outcome trait. The effect sizes of the five known mediators (M_1, M_2, M_3, M_4, M_5) on the outcome trait are (0.4, 0.2, 0.3, 0.2, 0.4). The known mediators are generated as Eqs. 13–17. The proportion of the SNPs that are associated with the hidden mediator, π_H, is 0.8. The hidden mediator is generated as Eq. 18. We also include two covariates, C_1 and C_2, in the outcome model. C_1 is generated from a normal distribution with a mean of 7 and a standard deviation of 0.5; C_2 is generated from a normal distribution with a mean of 4 and a standard deviation of 0.4. Their corresponding effect sizes on the outcome trait are 0.8 and -0.3. The outcome trait is generated based on Eq. 19. Note that we only add a relatively small error term (ε) to the outcome trait because we assume that almost all the leftover genetic heritability of the outcome trait can be explained by the hidden mediator. The mediators in the base case do not affect each other (i.e., no correlation among M_1, M_2, M_3, M_4, and M_5).

$$a_1 \sim \begin{cases} 0 \textit{ with probability } 0.5 \\ N\left(0.2,\ 0.08^2\right) \textit{ with probability } 0.5 \end{cases} \tag{7}$$

$$a_2 \sim \begin{cases} 0 \textit{ with probability } 0.4 \\ N\left(0.2,\ 0.08^2\right) \textit{with probability } 0.6 \end{cases} \tag{8}$$

$$a_3 \sim \begin{cases} 0 \textit{ with probability } 0.2 \\ N\left(0.2,\ 0.08^2\right) \textit{with probability } 0.8 \end{cases} \tag{9}$$

$$a_4 \sim \begin{cases} 0 \textit{ with probability } 0.8 \\ N\left(0.2,\ 0.08^2\right) \textit{with probability } 0.2 \end{cases} \tag{10}$$

$$a_5 \sim \begin{cases} 0 \textit{ with probability } 0.5 \\ N\left(0.2,\ 0.08^2\right) \textit{with probability } 0.5 \end{cases} \tag{11}$$

$$a_H \sim \begin{cases} 0 \textit{ with probability } 0.2 \\ N\left(0.2,\ 0.08^2\right) \textit{with probability } 0.8 \end{cases} \tag{12}$$

$$M_1 = 50 + \boldsymbol{a_1 G} + \varepsilon_1, \varepsilon_1 \sim N\left(0, 1^2\right) \tag{13}$$

$$M_2 = 5 + \boldsymbol{a}_2\boldsymbol{G} + \varepsilon_2, \varepsilon_2 \sim N\left(0, 1^2\right) \tag{14}$$

$$M_3 = 10 + \boldsymbol{a}_3\boldsymbol{G} + \varepsilon_3, \varepsilon_3 \sim N\left(0, 1.5^2\right) \tag{15}$$

$$M_4 = 6 + \boldsymbol{a}_4\boldsymbol{G} + \varepsilon_4, \varepsilon_4 \sim N\left(0, 1.2^2\right) \tag{16}$$

$$M_5 = 15 + \boldsymbol{a}_5\boldsymbol{G} + \varepsilon_5, \varepsilon_5 \sim N\left(0, 1^2\right) \tag{17}$$

$$M_H = 20 + \boldsymbol{a}_H\boldsymbol{G} + \varepsilon_H, \varepsilon_H \sim N\left(0, 1^2\right) \tag{18}$$

$$Y = 0.4M_1 + 0.2M_2 + 0.3M_3 + 0.2M_4 + 0.4M_5 + b_H M_H + 0.8C_1 - 0.3C_2 + \varepsilon, \varepsilon \sim N\left(0, 0.2^2\right) \tag{19}$$

Data application

We apply the proposed method to UK Biobank data. We consider Waist-to-hip Ratio (WHR) as the outcome, the significant SNPs from the latest GWAS meta-analysis of WHR as the exposure, and body mass index (BMI) as the known causal mediator between WHR and the SNPs [14]. The goal of the analysis is to determine whether there exists a second, hidden mediator on WHR and if so, estimate the effect size of this hidden mediator. In addition to waist circumference, hip circumference and BMI, we also include sex, age, and the first ten genetic principal components in the UK Biobank phenotype data as covariates in the mediation model. Only individuals with European ancestry are considered for the current analysis, which includes those described as "British", "Irish", "White" or "Any other white background". Individuals with missing phenotype data (i.e., "NA") in any of the data fields are removed.

Because the GWAS meta-analysis used for identifying the exposure SNPs involves UK Biobank data, to minimize over-estimated SNP effect sizes in our analyses, we use a stringent p-value threshold of 5×10^{-9} when choosing the exposure SNPs. The resulting SNPs are clumped using the *R* package *TwoSampleMR* (version 0.5.6) with a clumping window of 250 kb and a cutoff for correlation due to linkage disequilibrium (LD $r^2 = 0.01$) based on the 1000 Genomes Continental European groups reference [15, 16]. After LD clumping, a total of 535 independent SNPs associated with WHR are identified and extracted from the imputed UK Biobank genotype using *plink* (version 2.0) [17] (Additional file 2: Table S1). Furthermore, ten SNPs with low imputation quality (INFO score < 0.9) are dropped from further analyses (labeled with "INFO" in Additional file 2: Table S1), and a hard-call threshold of 0.4 is used when converting the imputed alleles probabilities to the number of allele copies. After joining the genotype data with the phenotype data, we further remove two SNPs with more than 10,000 missing rows (labeled with "NA" in Additional file 2: Table S1) and a multi-allelic SNP (labeled with "M" in Additional file 2: Table S1). Individuals with missing data (i.e., "NA") in any of the

data SNP fields are removed. In the end, 218,277 individuals and 522 SNPs are included in the application of our method. WHR is calculated as the ratio of waist circumference to hip circumference, and WHR and BMI are standardized to have means of zero and standard deviations of one.

The cleaned data are analyzed using the proposed multi-step method according to the flowchart in Fig. 2. In Step 1, an initial regression was performed to access the effect direction of the SNPs on the outcome trait WHR using WHR as the dependent variable and all the SNPs, sex, age, and the first 10 genetic principal components as the independent variables. Based on the direction (positive or negative) of the estimated effect, the coding of each SNP is flipped such that all SNPs have a positive effect on WHR. In the subsequent mediator (BMI) regression models (to obtain $\boldsymbol{a}^*$) and the outcome (WHR) regression model (to obtain $\boldsymbol{c}^*$), sex, age and the first ten principal components are adjusted for as covariates. For both $\boldsymbol{a}^*$ and $\boldsymbol{c}^*$, SNPs with effects that are greater than the 3rd quartile $+ 3 \times$ interquartile range (IQR) and values that are smaller than the 1st quartile—$3 \times$ IQR are removed to avoid the downstream GMM methods to be driven by these outliers. A total of 12 SNPs were removed (label with "O" in Additional file 2: Table S1). In Step 3, the MCMC chain length is set to 30,000 with a burn-in length of 5,000. We also estimate BMI's effect on WHR by fitting a regression model with the dependent variable being WHR and the independent variables being BMI, sex, age and the first ten principal components. As a sensitivity analysis, we repeated the proposed multi-step method multiple times with some SNPs dropped. Specifically, we randomly divided 522 SNPs into 20 groups (i.e., approximately 26 SNPs per group), and the analysis was repeated 20 times with each of the 20 groups dropped.

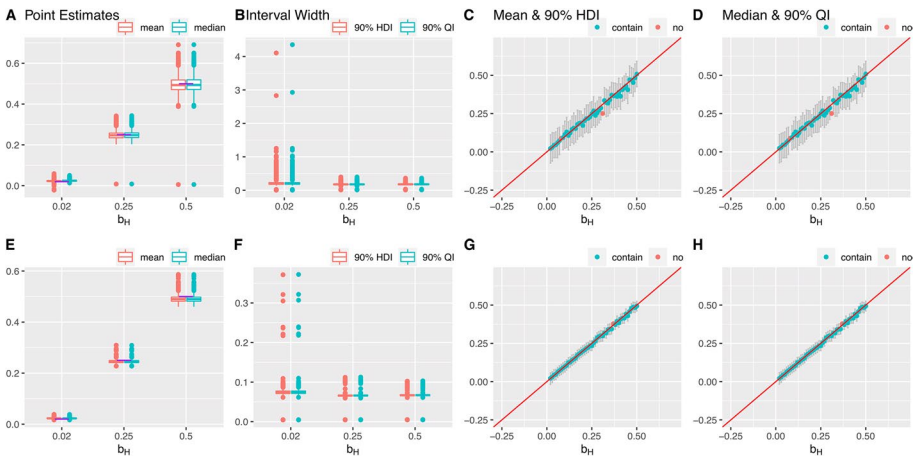

Fig. 3 Results of one simulation setting: base case, varying b_H, 1000 simulations. The first and second row presents the results for 70 SNPs and 500 SNPs, respectively. (**A**, **E**) Box plots of the posterior median and the mean of b_H. The purple lines indicate the true values. (**B**, **F**) Box plots of the widths of 90% HDIs and QIs. (**C**, **G**) The posterior medians and the 90% HDIs of the 49 equally spaced values of b_H between 0.02 and 0.5. (**D**, **H**) The posterior means and the 90% QIs of the 49 equally spaced values of b_H between 0.02 and 0.5. Outliers are defined as the values more extreme than the third quartile + 1.5 * (the third quartile—the first quartile) or the first quartile—1.5 * (the third quartile—the first quartile)

Results

Simulation results

For the base case, simulation results on the median, mean, 90% highest density interval (HDI) and 90% quantile interval (QI) of b_H's posterior distributions are summarized in Fig. 3. Rather than the usual 95% intervals, 90% intervals are reported to avoid the potential instability at the tails of the MCMC posterior distributions. We also report the root mean square error, the average bias, the number of outliers of the mean and the median point estimators, the proportion of the times that the HDI and QI contain the true value of b_H, and empirical power / type I error of the HDI and the QI interval estimators for each of the eight simulation settings (Additional files 3 and 4: Tables S2 and S3). The outliers are defined as the values more extreme than the third quartile + 1.5 * (the third quartile − the first quartile) or the first quartile − 1.5 * (the third quartile − the first quartile). For the base case and Settings 1–6, the empirical power is calculated as the number of simulations with the interval not containing zero divided by the total number of simulations. For Setting 7, the empirical type I error is calculated as the number of simulations with the interval not containing zero divided by the number of total simulations.

Under the base case, the performances of the posterior median and mean of b_H are similar; both are close to the true value of b_H, and slightly downward biased only when the true value of b_H is 0.5. Both the median and the mean have a smaller variation for the cases with 500 SNPs compared to 70 SNPs (Fig. 3A –H). In the case with 70 SNPs, where both $b_H = 0.25$ and $b_H = 0.5$, for one out of the one thousand simulation runs the median and the mean were far away from the corresponding true value (at values close to zero) (Fig. 3A). But we did not observe this in the 500 SNPs case. The 90% HDI and the 90% QI behaved similarly as well; the intervals were wider for the cases with 70 SNPs than 500 SNPs only a few times, and the HDI and the QI were extremely wide. For $b_H = 0.02$, $b_H = 0.25$ and $b_H = 0.5$ 98.1%, 98.3%, and 97.3% of the 90% HDI and the 90% QI contains the true value of b_H, indicating both intervals were conservative. In addition, when the true value of b_H is small, both interval estimators were slightly wider. Although the behavior of the posterior distribution of b_H did not change dramatically with the number of SNPs in the model, having more SNPs in the model can lead to slightly better estimations of b_H in terms of both the point estimators and the interval estimators.

Simulations were also conducted for eight additional settings. In Setting 1, where a low proportion of the SNPs are associated with the hidden mediator, the posterior median and meanindicated slightly larger downward biases, and the HDI and the QI were wider when there is a smaller number of SNPs in the model. In contrast, when all of the SNPs were associated with the hidden mediators, the posterior median and mean were biased upward and the HDI and QI were less likely to capture the true value of b_H when the true value of b_H is large (Additional file 1: Sect. 3.1). For Settings 2, 3 and 4, the simulation results showed that having known mediators with negative effects on the outcome trait and varying the sample size (number of individuals) and the number of known mediators did not have dramatic impacts on the posterior distribution of b_H (Additional file 1: Sect. 3.2, 3.3, 3.4). We also observe that in Setting 5, the posterior median and mean can be biased and the HDI and the QI are less likely to include the true value of b_H if the causal relationships among the known mediators are not appropriately adjusted for

in the mediation regressions during Step 1 (Additional file 1: Sect. 3.5). Furthermore, if the assumption that the SNP effects on all the mediators come from the same distribution does not hold, depending on the degree, the posterior median and mean can vary greatly and the HDI and the QI can have a low chance to include the true value of b_H as shown in Setting 6 (Additional file 1: Sect. 3.6). Next, according to the simulation result for Setting 7, the posterior median and meanwere close to the true value of b_H, zero, and the HDI and the QI have Type I error rates thatwere close to 0.1 (Additional files 1 and 4: Sect. 3.7 and Table S3). Finally, the simulation result for Setting 8 showed that as the number of SNPs in the model decreases, the interval estimates become wider, and the point estimators become less precise, which is as expected. However, the point and interval estimators still have decent performance when therewere only 20 SNPs in the model, indicating that the performance of our method is not greatly affected as long as there are asufficient number of SNPs in the model. (Additional file 1 and 4: Sect. 3.8, Table S3). Detailed results of the additional settings are presented in the Additional files.

Application on waist-hip ratio

The regression estimated SNP effects on BMI ($\boldsymbol{a}^*$) and direct effects between the SNPs and WHR $\boldsymbol{c}^*$ are shown in Fig. 4A and C, respectively. As shown in the histograms, therewere some extreme values or outliers in both $\boldsymbol{a}^*$ and $\boldsymbol{c}^*$. The histograms with

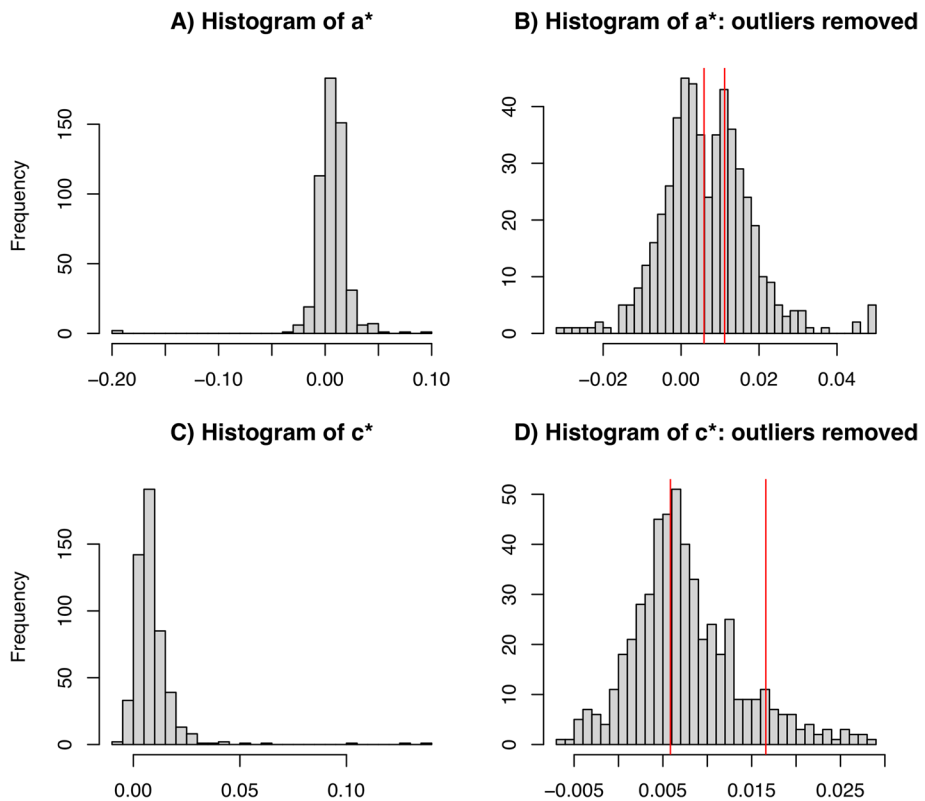

Fig. 4 Histograms of the regression estimated effects. (**A**) Estimated SNP effects on BMI (a^*). (**B**) Estimated SNP effects on BMI with outliers removed. (**C**) Estimated direct effects between the SNPs and WHR (c^*). (**D**) Estimated direct effects between the SNPs and WHR with outliers removed

outliers removed are shown in Fig. 4B and D. Two modes can be observed from the distribution of $\boldsymbol{a}^*$ with the mode on the left being approximately at zero suggesting $\boldsymbol{a}^*$ follow a mixture model. It is less clear whether the distribution of $\boldsymbol{c}^*$ has two modes. EM estimated means of $\boldsymbol{a}^*$ and $\boldsymbol{c}^*$ are indicated by the vertical red lines shown in Fig. 4B and D. The fitted GMM on $\boldsymbol{a}^*$ estimates a mixture weight of 0.882 for the distribution on the left, and the fitted GMM on $\boldsymbol{c}^*$ estimates a mixture weight of 0.836 to the distribution on the left. From the posterior distribution of b_H, the point estimate for b_H using the posterior median is 1.5556 and 1.5562 using the posterior mean; the HDI is (1.4658, 1.6440); the QI is (1.4683, 1.6469). The whole procedure of the multi-step method on waist-hip-ratio (218,277 individuals, 522 SNPs, 1 known mediator, 5,000 iterations for the MCMC burn-in, and 30,000 iterations for the MCMC chain length) took 6.33 min on a MacBook Pro laptop with a 2.8 GHz Intel Core i7 processor and 16 GB memory. The runtime can be variable based on how long the MCMC chain is set to. In the sensitivity analysis that repeats the multi-step method 20 times with approximately 1/20 of the SNPs dropped each time, the median of the posterior median of b_H from 20 repeats is 1.5486, the first quartile is 1.4483, the third quartile is 1.6355, the smallest value is 1.0511, and the largest value is 1.9575. For the posterior mean of b_H, the median of the 20 repeats is 1.5494, the first quartile is 1.4492, the third quartile is 1.6368, the smallest value is 1.0511, and the largest value is 1.9584. The posterior median and mean from three repeats are relatively extreme (posterior median = 1.0511, 1.0643, 1.9574) indicating that some SNPs are more influential than other SNPs. Finally, the effect of BMI on WHR conditional on sex, age and the first ten principal components is estimated to be 0.3834 (95% CI: 0.3807, 0.3861).

Discussion

In this work, we propose to infer the effect size of a potential hidden mediator on a trait of interest based on the observed associations between the trait of interest, its known causal factors, and the associated SNPs that have been identified previously. Utilizing the mediation framework, we propose a multi-step method to estimate the effect size of the hidden factor by treating the trait of interest, its associated SNPs, and the known causal factors as the outcome, exposure, and known mediators in the mediation model, respectively. Assuming the direct effects between the outcome trait and the identified SNPs that are unexplained by the known mediators can be explained by a hidden mediator, we obtain the effect of this hidden mediator on the outcome trait by decomposing the direct effects between the outcome trait and its associated SNPs into the SNP effects on the hidden mediator and the hidden mediator's effect size on the outcome trait. In Step 1 of our proposed method, we estimate the SNP effects on the known mediators and the direct effects between the SNPs and the outcome trait via a series of linear regressions. In Step 2, we fit two GMMs on the estimated SNP effects on the known mediators and the direct effects between the SNPs and the outcome trait using the EM algorithm to separate the actual SNP effects from the zero-mean noises (i.e., the estimated effects of those SNPs not associated with the known and hidden mediators). Lastly in Step 3, based on the EM estimated GMM parameters, a GMM MCMC procedure is applied to generate a posterior distribution for the hidden mediator's effect size, b_H.

Through extensive simulation studies, we show that our method can produce a posterior distribution that captures b_H well. We observe that if the model assumptions are

correct, both using the posterior median and mean to estimate b_H provides only small biases, and good coverage of the simulated true effect by the 90% HDI and QI. When there are more identified SNPs associated with the outcome trait, in general, both the point and the interval estimators perform better. Also, our posterior distribution estimates the hidden mediator's effect size well if some of the known mediators have negative effects on the outcome trait while others have positive effects. Our method can also accommodate varying numbers of known mediators and the performance is not dramatically affected by decreasing sample size. In addition, our method estimates b_H well even if the hidden mediator does not exist (i.e., $b_H = 0$). However, we notice that when the hidden mediator is associated with all of the identified SNPs that are associated with the outcome trait, both the median and the mean are upward biased and both the HDI and the QI are less likely to include the true value of b_H. This is expected because the assumed GMM distribution on the estimated direct effects between the SNPs and the outcome trait is wrong. On the other hand, if the hidden mediator is associated with too few identified SNPs, the point estimators have large biases and the interval estimators become wider. This is also expected as there are fewer SNPs that can provide information about b_H. Also, if causal relationships exist among the known mediators, these effects need to be adjusted in the mediation regressions. Otherwise, the point estimators can have relatively large biases and the interval estimators are likely to include the true value if the causal relationships are large enough. Finally, our method assumes that the SNP effects on different mediators (including both the known and hidden mediators) are similar. The posterior distribution estimates b_H poorly when the SNP effects on different mediators are very different, especially when the number of SNPs in the model is large. This is because little information can be borrowed from the observed SNP effects to infer the SNP effects on the hidden mediator, which can make it difficult to decompose the direct effects between the outcome trait and the associated SNPs to estimate b_H.

Occasionally, a posterior distribution with little information about b_H (i.e., is extremely wide) can be generated when the true value of b_H is either zero or very close to zero such that the distribution of the estimated direct effects between the outcome trait and the associated SNPs follows a single Gaussian distribution around zero rather than a GMM. Under this scenario, by chance alone, the EM algorithm in Step 2 may assign a tiny weight to the distribution that involves b_H in the MCMC model in Step 3 and assign a large weight to the other distribution in the GMM. As a result, little data can be used to infer b_H in the MCMC procedure. To avoid this situation, we suggest one inspect the histogram of estimated direct effects and the EM fitted GMM from Step 2. If the histogram of estimated direct effects does not have two modes and is centered approximately at zero, and the EM fitted GMM assigns a tiny weight to the distribution that involves b_H relative to the weight of the other distribution, then it is reasonable to flip the binary labeling of GMM and proceed to Step 3.

We applied our proposed method on UK Biobank data to estimate the effect size for a potential hidden mediator of waist-hip ratio. From the posterior distribution generated by our method, the posterior median estimates that a potential hidden mediator exists in the European population with an effect size of 1.56 (90% QI: 1.47, 1.64). This result suggests that the hidden mediator has a larger effect on waist-hip ratio comparing to BMI (0.38). Some caution here is warranted, as we used the same UK Biobank data for

both identifying SNPs associated with waist-hip ratio and for estimating the SNP effects. Although we used an extra stringent p-value threshold (5×10^{-9}) for filtering SNPs associated with waist-hip ratio to mitigate biases from winner's curse, some degree of biases from the winner's curse is unavoidable. A more optimal approach is to identify SNPs associated with waist-hip ratio and perform the estimation in two independent populations with similar ancestries such that the effects are similar in the two populations but the biases from the winner's curse are minimal. There are a couple of possible biological or physiological explanation for the hidden mediator identified in the current analysis that could contribute to the remaining association. For example, the distribution of brown fat (the cell type responsible for non-shivering thermogenesis) varies across individuals and could influence body size and weight, and studies have reported the negative association between WHR and having active brown fat in males [18] and the negative association between WHR and neuregulin 4, an adipokine secreted by brown fat, in children [19]. However, there are no great quantitative measurements of that trait nor have genetic studies of this trait been performed. Another possibility is the propensity for physical activities, which can be measured by actigraphy and could certainly influence body size and weight [20, 21].

Our proposed method has some limitations. First, the performance of the posterior distribution for b_H under our method largely depends on the how well we estimate the regression coefficients for the mediation regressions during Step 1, as the downstream steps treat the estimated regression coefficients as input data. Precise and accurate estimates of the coefficients require the mediation regression to be performed on data sets with large samples size, especially when many SNPs are included in the model. For large population genetics data, this may be less of a concern. As we learned from the simulation studies, the posterior distribution captures b_H well when the sample size is 25,000, which for today's conventional size of DNA biobanks is not unreasonable. However, if strong correlations among the known mediators are not properly adjusted in the mediation regressions, the regression coefficients will be biased, which can lead to substantial biases in the resulting posterior of b_H. Thus, our method relies on one's input of prior domain knowledge and the specification of reasonable regression models. Future work can be devoted to extending the methods to address the situation when the known mediators are correlated with each other and no prior knowledge on the causal directions among the known mediators is available. Second, our method relies on the strong assumption that the SNP effects on all the mediators between the SNPs and the outcome trait come from the same distribution, and departure from this assumption can lead to substantial variation for the posterior distribution for b_H such that the inference based on the posterior distribution can be very inaccurate. However, we argue that it is somewhat reasonable to constrain the SNP effects on the hidden mediator to be similar to the SNP effects on the known mediators and there will be infinite number of ways to decompose the direct effect between the SNPs and the outcome trait without this assumption. Future work can focus on relaxing this assumption or inventing ways to incorporate information about the potential SNP effect sizes from another modality. Furthermore, we view our work that considers the simplest case with one hidden mediator as an initial step in learning about the residual associations between an outcome and the existing mediator. The current model cannot distinguish whether the estimated effect size is for

one hidden mediator or the combined effect size of multiple hidden mediators. In reality, any leftover associations—and perhaps more likely—could be due to multiple hidden mediators each with a relatively small effect size. Even if there are multiple hidden mediators, our approach to determine whether the known mediators fully explain the heritability of the outcome trait is still applicable, and we can interpret the estimated effect size assuming a single hidden mediator as the combined effect of multiple potential hidden mediators. Future work can be devoted to developing a method capable of inferring the number of potential hidden mediators and decomposing the combined effect into individual effects from each of the hidden mediators. Furthermore, specifying priors for the SNP effect sizes based on the heritability model, as suggested by one of the reviewers, may be a way to incorporate the estimation uncertainties of the SNP effect sizes into the model, lead to a better estimation of the hidden mediator's effect size. Lastly, future work can extend the continuous outcome trait to other outcome types such as binary variables potentially via the counterfactual framework. This will involve utilizing generalized linear models in the meditation regressions in Step 1. Such extension will make the method useful for many applications such that the disease trait is binary.

Conclusions

We developed a method for estimating the effect size of a potential hidden estimator between a trait of interest and its associated SNPs. In the first step, a series of regression models are used to estimate the SNP effects on the hidden mediators and the direct effects between the SNPs and the trait of interest. In the second step, GMM models are fitted to the estimated SNP effects on the hidden mediators and the estimated direct effects between the SNPs via the EM algorithm. In the final step, an MCMC procedure that utilizes parameters estimated in the second step is used for generating a posterior distribution for the hidden mediator's effect size. Extensive simulations show that our method can generate accurate estimators for the hidden mediator's effect size. Also, when the hidden mediator does not exist, our method has controlled type I error rates. By applying our method to UK Biobank data, we found a potential hidden mediator between waist-hip-ratio and its associated SNPs in the European population and estimated its effect size on waist-hip-ratio to be larger than a known mediator BMI's effect size on waist-hip-ratio. Although, as an initial step toward finding the hidden mediators between a trait of interest and its associated SNPs, we hypothesize a simple model with only one hidden mediator left, we hope that our method can provide some insights into the characteristics of the remaining one or multiple hidden mediators and inspire further method developments in this direction.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04977-4.

Additional file 1. A pdf file containing supplementary text (Sections S1–S4) and figures (Figures S1–S20) including the discussion on a more generation assumption on the SNP effects (Section S1), additional simulation settings and results (Sections S2–S3), and a proof of the estimated SNP effect's distribution (Section S4).

Additional file 2. An Excel table containing the list of 535 independent SNPs associated with WHR.

Additional file 3. An Excel table containing the simulation results for Settings 1–6 and 8

Additional file 4. An Excel table containing the simulation results for Setting 7.

Acknowledgements
Not applicable.

Author contributions
Z.D. and B.F.V. conceived of the project and designed the experiments. M.D.R. provided data. Z.D. analyzed the data. Z.D., W-T.H., and B.F.V. developed the method and prepared a draft of the initial manuscript. All authors edited the manuscript. W-T.H and B.F.V. supervised the project. All authors read and approved the final manuscript.

Funding
B.F.V. acknowledges support from the National Institutes of Health (DK101478, DK126194) and a Linda Pechenik Montague Investigator Award. M.D.R. acknowledges support from the National Institutes of Health (AI077505).

Availability of data and materials
The UK Biobank data are provided under application ID 32133 and can be accessed by others at the UK Biobank website (https://www.ukbiobank.ac.uk). R Scripts with the simulation code and an R package that implements the described method are available at https://github.com/zhd007/HiddenMediator [22].

Declarations

Ethics approval and consent to participate
We confirm that all methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors have no conflicts to report.

Received: 8 June 2022 Accepted: 6 October 2022
Published online: 13 October 2022

References

1. Thom CS, Ding Z, Levin MG, Damrauer SM, Lee KM, Lynch J, Chang KM, Tsao PS, Cho K, Wilson P, Assimes TL, Sun YV, O'Donnell CJ, Million Veteran Program VA, Vujkovic M, Voight BF. Genetic determinants of increased body mass index mediate the effect of smoking on increased risk for type 2 diabetes but not coronary artery disease. Hum Mol Genet. 2020;29(19):3327–37.
2. Maddatu J, Anderson-Baucum E, Evans-Molina C. Smoking and the risk of type 2 diabetes. Transl Res. 2017;184:101–7.
3. Herder C, Karakas M, Koenig W. Biomarkers for the prediction of type 2 diabetes and cardiovascular disease. Clin Pharmacol Ther. 2011;90(1):52–66.
4. Njajou OT, Kanaya AM, Holvoet P, Connelly S, Strotmeyer ES, Harris TB, Hsueh WC. Association between oxidized LDL, obesity and type 2 diabetes in a population-based cohort, the health, aging and body composition study. Diabetes Metab Res Rev. 2009;25(8):733–9.
5. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annu Rev Psychol. 2007;58:593.
6. VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. Epidemiol Methods. 2014;2(1):95–115.
7. Huong PT, Nguyen CTT, Nhung VT. The association between FTO polymorphisms and type 2 diabetes in Asian populations: a meta-analysis. Meta Gene. 2021;30: 100958.
8. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
9. Peel DA, MacLahlan G. Finite mixture models. John & Sons. 2000.
10. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika. 1993;80(2):267–78.
11. Tatiana B, Didier C, David RH, Derek Y. mixtools: an R package for analyzing finite mixture models. J Stat Softw. 2009;32(6):1–29.
12. Plummer M, Stukalov A, Denwood M. Package rjags: Bayesian graphical models using MCMC. R package version. 2016:4-6.
13. Plummer M (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, No. 125.10, pp. 1–10).
14. Pulit SL, Stoneman C, Morris AP, Wood AR, Glastonbury CA, Tyrrell J, Lindgren CM. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum Mol Genet. 2019;28(1):166–74.
15. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY. The MR-Base platform supports systematic causal inference across the human phenome. elife. 2018;7.
16. Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68.

17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
18. Zhang Q, Ye H, Miao Q, Zhang Z, Wang Y, Zhu X, Li Y. Differences in the metabolic status of healthy adults with and without active brown adipose tissue. Wien Klin Wochenschr. 2013;125(21):687–95.
19. Wang R, Yang F, Qing L, Huang R, Liu Q, Li X. Decreased serum neuregulin 4 levels associated with non-alcoholic fatty liver disease in children with obesity. Clin Obes. 2019;9(1): e12289.
20. Doherty A, Smith-Byrne K, Ferreira T, Holmes MV, Holmes C, Pulit SL, Lindgren CM. GWAS identifies 14 loci for device-measured physical activity and sleep duration. Nat Commun. 2018;9(1):1–8.
21. Klimentidis YC, Raichlen DA, Bea J, Garcia DO, Wineinger NE, Mandarino LJ, Going SB. Genome-wide association study of habitual physical activity in over 377,000 UK Biobank participants identifies multiple variants including CADM2 and APOE. Int J Obes. 2018;42(6):1161–76.
22. HiddenMediator. https://github.com/zhd007/HiddenMediator. Accessed 4 June 2022.

Publisher's Note