frontiers in
GENETICS

METHODS ARTICLE
published: 21 November 2013
doi: 10.3389/fgene.2013.00243

Testing for direct genetic effects using a screening step in family-based association studies

Sharon M. Lutz[1,2]*, Stijn Vansteelandt[3] and Christoph Lange[2]

[1] Department of Biostatistics, University of Colorado, Aurora, CO, USA
[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

Edited by:
Xiangqing Sun, Case Western Reserve University, USA

Reviewed by:
Nathan Morris, Case Western Reserve University, USA
Sungho Won, Chung-Ang University, South Korea

***Correspondence:**
Sharon M. Lutz, Department of Biostatistics, University of Colorado, 13001 E. 17th Place, B119 Building 500, Room W3128, Anschutz Medical Campus, Aurora, CO 80045, USA
e-mail: sharon.lutz@ucdenver.edu

In genome wide association studies (GWAS), family-based studies tend to have less power to detect genetic associations than population-based studies, such as case-control studies. This can be an issue when testing if genes in a family-based GWAS have a direct effect on the phenotype of interest over and above their possible indirect effect through a secondary phenotype. When multiple SNPs are tested for a direct effect in the family-based study, a screening step can be used to minimize the burden of multiple comparisons in the causal analysis. We propose a 2-stage screening step that can be incorporated into the family-based association test (FBAT) approach similar to the conditional mean model approach in the Van Steen-algorithm (Van Steen et al., 2005). Simulations demonstrate that the type 1 error is preserved and this method is advantageous when multiple markers are tested. This method is illustrated by an application to the Framingham Heart Study.

Keywords: family-based association analysis, causal inference, genetic pathway, mediation, pleiotropy

INTRODUCTION

Some of the recently published genome-wide association studies identified the same genetic locus as a disease susceptibility locus for different complex diseases (Amos et al., 2008; Thorgeirsson et al., 2008). One possible mechanism is that the marker locus is pleiotropic and has genetic effects on several, different phenotypes. Determining whether the marker acts directly on each of these phenotypes or only indirectly via one or more intermediate phenotypes is an important step in understanding the biological significance of the genetic associations. In order to understand and characterize the underlying genetic effect, methods have been proposed to disentangle these potential direct and indirect genetic effects (Vansteelandt et al., 2009; Vansteelandt, 2010; Berzuini et al., 2012; Vansteelandt and Lange, 2012; VanderWeele et al., 2012). All currently available methods focus on the direct and indirect genetic effects relative to one (group of) secondary phenotypes. Because the magnitude of the indirect effect depends on how strongly these secondary phenotypes affect the primary phenotype, these methods consider adjustment for confounding of the relationship between these phenotypes by measured extraneous factors. Some of these methods quantify both the direct and indirect genetic effects, but assume that none of these extraneous confounding factors is influenced by the considered marker (VanderWeele et al., 2012). Some of these methods allow for some of the extraneous confounding factors to be influenced by the considered marker, but they merely quantify direct genetic effects (Vansteelandt et al., 2009; Vansteelandt, 2010; Berzuini et al., 2012).

Regardless of the considered framework, all available methods only test one gene at a time and need to be corrected for multiple comparisons. This concern over multiple comparisons becomes an issue in family-based genome wide association studies (GWAS). When there is a region with a strong association with both the endo-phenotype and phenotype, identifying SNPs in the region that are still associated with the phenotype of interest after accounting for the association with the endo-phenotype requires testing for a direct causal effect for every SNP in the region. In order to increase power to detect this direct genetic effect, we propose a 2-stage testing strategy to minimize the burden of multiple comparisons in the causal analysis (Van Steen et al., 2005; Murphy et al., 2008; Won et al., 2009). The application of a screening step when testing for direct genetic effects is an important advantage in this scenario where the multiple-comparison problem is a major hurdle. The power of our approach is assessed by simulation studies. We show that the type-1 error is preserved and the method is shown to be advantageous when multiple SNPs are tested for a direct effect on the phenotype of interest.

METHODS

Suppose that in the family-based study, n trios (offspring and both parents) have been genotyped at a specific marker locus. Assuming there is no bias due to ascertainment conditions, the variable X_i denotes the coded genotype of the offspring and S_i denotes the parental genotypes for individual i. If genotypic data is unavailable for the parents but genotypic information is available on the subject's siblings, the variable S_i denotes the sufficient statistic by Rabinowitz and Laird (2000) For offspring i, Y_i denotes the target phenotype in the association study and K_i denotes the secondary phenotype in the study.

Suppose that an association has been observed between the secondary phenotype of interest, K_i, and the marker locus. Given this association, the goal is to test for an association between the target phenotype Y_i and the marker locus that cannot be explained by a possible indirect effect mediated by K_i. To achieve this goal, data is needed on all risk factors of the secondary phenotype K_i that are also associated with the primary phenotype (Cole and Hernan, 2002). Let L_i denote this collection of measured confounding variables. Because L may be high-dimensional, we do not assume that it is only related with Y by means of a causal effect, but allow for their association to be itself confounded by potentially unmeasured factors U. This is shown in the causal diagram of **Figure 1**, where the presence of U additionally captures the potential for confounding of the genetic association as a result of population admixture (Vansteelandt and Lange, 2012). Throughout, in contrast to other mediation analysis techniques (namely those based on so-called natural direct and indirect effects), we will allow for the possibility that some of these confounding variables are themselves affected by the studied marker, as illustrated via the edge from X to L in the causal diagram (VanderWeele et al., 2012).

Consider model

$$E[Y_i|X_i, K_i, L_i] = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 L_i \tag{1}$$

where γ_j for $j = 0, 1, \ldots, 3$ denote the mean parameters and can be estimated by ordinary least squares. Note that γ_1 represents the true effect of K_i on Y_i and not a spurious association because, by assumption, the above model includes all relevant risk factors of K_i. In order to construct an adjustment principle that tests for a direct genetic effect of the marker locus X on the target phenotype Y, the effect of the secondary phenotype K has to be estimated. Vansteelandt et al. use an estimate for γ_1 based on model (1) to adjust the phenotype Y_i to $Y_i - \gamma_1 K_i$. A family-based association test (FBAT) on this adjusted phenotype is then a test for the direct genetic effect in the family-based setting (provided that the distribution of the test statistic acknowledges the uncertainty in the estimated effect γ_1) (Vansteelandt et al., 2009).

To reduce the number of multiple comparisons, we adapt the conditional mean model approach in the VanSteen-algorithm (Van Steen et al., 2005) to model (1). By replacing the observed marker score in model (1) by the expected marker score conditional upon the parental genotypes or sufficient statistic, the genetic effects of locus X_i can be assessed without having to adjust the α-level of any subsequently computed FBATs (Lange et al., 2003a,b; Van Steen et al., 2005). Similar to the idea of the conditional mean model approach, model (1) can be rewritten by substituting X_i with its expected value $E(X_i|S_i)$,

$$E[Y_i|K_i, L_i, S_i] = \beta_0 + \gamma_1 K_i + \beta_2 L_i + \beta_3 E(X_i|S_i)\,, \tag{2}$$

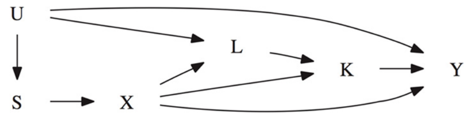

FIGURE 1 | Causal diagram illustrating the confounding of the target phenotype *Y* and the marker locus *X*. *S* denotes the parental genotype or Rabinowitz and Laird's sufficient statistic. *K* denotes the secondary phenotype of interest. *L* allows for confounding between *K* and *Y*. *U* represents a collection of unmeasured factors that allow for confounding due to population stratification or confounding between the two phenotypes *K* and *Y*. Note that causal diagrams assume that all variables that jointly affect any two variables are included. The absence of an arrow between any two variable denotes that there is no direct causal effect. For instance, *U* has no direct causal effect on *X*.

As shown in the proof given in the appendix, the parameter γ_1 is the same in both model (1) and model (2) when the null hypothesis holds that there is no direct effect and, moreover, there is no confounding due to population substructure. For testing the null hypothesis of no direct genetic effect, model (2) can thus be used to estimate the parameter γ_1 in a screening step without biasing the significance level since X_i is not included in this model, provided there is no confounding due to population substructure.

For the screening step, each subject contributes

$$T_i^* = \{E(X_i|S_i)\}\,\tilde{Y}_i^* \tag{3}$$

where $\tilde{Y}_i^* = Y_i - \bar{y} - \hat{\gamma}_1^*(K_i - \bar{k})$ and $\hat{\gamma}_1^*$ is the ordinary least squares estimate for γ_1 in model (2), which does not involve the genetic marker X. $\tilde{Y}_i^*$ is not adjusted for the covariates L_i since including factors such as L_i in the phenotypic adjustment would introduce bias if the common risk factor L_i is associated with the DSL X_i (Vansteelandt et al., 2009). The parameters $\bar{y}$ and $\bar{k}$ are the observed phenotypic means of Y and K in the sample, respectively. Then the test statistic for the screening step is

$$\frac{\left(\sum_{i=1}^{n} T_i^*\right)^2}{\sum_{i=1}^{n} \mathrm{var}(\tilde{T}_i^*)} \tag{4}$$

where

$$\tilde{T}_i^* = T_i^* - E\big[\{E(X_i|S_i)\}\,(K_i - \bar{k})\big]\,\frac{\left(K_i - \mu_k^{*(i)}\right)}{\sigma_k^{*2}}\,\epsilon_i^* \tag{5}$$

where $\mathrm{var}(\tilde{T}_i^*)$ is calculated based on the sample variance of $\tilde{T}^*$ and ϵ_i^* denotes the residual from model (2). $\mu_k^{*(i)} = E(K|L_i, E(X_i|S_i))$ is the predicted value for K under a linear regression model for K with the covariates L_i and $E(X_i|S_i)$, and σ_k^{*2} denotes the residual variance in that model. The variance correction given in Equation (5) is needed to account for estimating γ_1 in the proposed phenotype adjustment $\tilde{Y}_i^*$ (Vansteelandt et al., 2009).

For step 1, the test statistic given in Equation (4) can be used for the screening step to pick the SNPs with the highest power since X is not used in this test statistic. For step 2, this smaller subset of SNPs are used to test the null hypothesis of no direct effect using the test statistic based on Equation (1) proposed by Vansteelandt et al. (2009)

$$T_i = \{X_i - E(X_i|S_i)\}\tilde{Y}_i \tag{6}$$

where $\tilde{Y}_i = Y_i - \bar{y} - \hat{\gamma}_1(K_i - \bar{k})$ and $\hat{\gamma}_1$ is the ordinary least square estimate for γ_1 in model (1), which does involve the genetic marker X. Using this association test with the adjusted phenotype $\tilde{Y}_i$ as the target phenotype provides a robust and valid test for the null hypothesis that there is no direct effect between the target phenotype Y_i and the DSL; i.e., the association between the target phenotype Y_i and the DSL is solely a result of the association between the secondary phenotype K_i and the DSL. Adjusting for estimating γ_1 based on model (1), the test statistic is distributed chi-square with one degree of freedom under the null hypothesis of no direct effect of X on Y and has the following form

$$\frac{\left(\sum_{i=1}^{n} T_i\right)^2}{\sum_{i=1}^{n} \text{var}(\tilde{T}_i)} \tag{7}$$

where

$$\tilde{T}_i = T_i - E[\{X_i - E(X_i|S_i)\}\, K_i]\, \frac{\left(K_i - \mu_k^{(i)}\right)}{\sigma_k^2}\epsilon_i \tag{8}$$

where $\text{var}(\tilde{T_i})$ is calculated based on the sample variance of $\tilde{T}$ and ϵ_i denotes the residual from model (1). $\mu_k^{(i)} = E(K|L_i, X_i, E(X_i|S_i))$ is the predicted value for K under a linear regression model for K with the covariates L_i, X_i, and $E(X_i|S_i)$, and σ_k^2 denotes the residual variance in that model. The variance correction given in Equation (8) is needed to account for estimating γ_1 in the proposed phenotype adjustment $\tilde{Y}_i$ (Vansteelandt et al., 2009). Note that Equation (3) is similar to Equation (6), but Equation (6) contains the genetic marker X_i. Similarly, Equation (5) is similar to Equation (8), but Equation (8) contains the genetic marker X_i.

Note that under the alternative hypothesis, the association between K and Y is different in models (1) and (2), even in the absence of population admixture. Model (1) represents the causal effect of K on Y under the alternative hypothesis, but model (2) does not represent the causal effect of K on Y because there is a remaining spurious association between X and Y along the path $K \leftarrow X \rightarrow Y$ in **Figure 1**. Under the null hypothesis, this path does not exist. As a result, the proposed approach is valid for testing in the absence of population stratification, but may have less power when either the $X \rightarrow K$ or the $X \rightarrow Y$ link is strong. This scenario is explored further in the simulation section of this paper.

Because the test statistic for the screening step given in Equation (4) is susceptible to population stratification, we examined this scenario in the simulation section as well. Principal component analysis (PCA) can be used in the screening step to correct for population stratification.

SIMULATIONS

Using simulation studies, we asses the type-1 error rate, the power, and robustness of this new approach which uses a trait that estimates γ_1 based on model (2) in the screening step and compare it to the approach proposed by Vansteelandt et al. (2009) which uses a trait that estimates γ_1 based on model (1). Similar to Vansteelandt et al. (2009), both methods are evaluated under various conditions. All simulations use a sample size of 1000 trios and are based on 5000 replications. The simulations are run for allele frequencies 5, 10, 15, 20, 25, 30, 35, 40, and 45%.

To reflect a realistic setting, the data is simulated to reflect covariances found in the Framingham Heart Study (Herbert et al., 2006). The phenotype of interest Y is simulated such that it resembles FEV1. The secondary phenotype K resembles weight and the set of common confounding variables resemble height and age. As seen in **Figure 2**, the first scenario assumes there is a direct genetic effect of the marker on the intermediate phenotype K and on the common covariate L. Each genetic effect has a locus specific heritability of 1%. The intermediate phenotype K explains 1% of the phenotypic variation in Y, creating an association between the SNP and Y. The second scenario is similar to the first scenario except that there is no genetic effect on the confounder L. The genetic association with the intermediate phenotype K is still present. The third scenario is similar to the first scenario except there is no association between K and Y. The fourth scenario is similar to the second scenario except that there is no genetic effect on the intermediate phenotype K.

As seen in **Table 1**, the type-1 error rate is similar whether model (1) or model (2) is used to estimate γ_1. For lower allele frequencies, under scenario 1 and 3, the type-1 error rate is 1–2% higher than expected. For higher allele frequencies under all four scenarios, the type-1 error rate is 0.5% lower than expected. In general, the type-1 error rate is close to 0.05 regardless of how γ_1 is estimated. As seen in **Table 2**, the power is similar whether model (1) or model (2) is used to estimate γ_1 assuming no population admixture. For lower allele frequencies, the method by

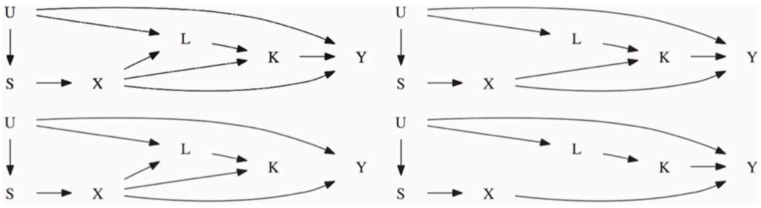

FIGURE 2 | The top left figure represents scenario 1. The top right figure represents scenario 2 which is the same as scenario 1 except that X does not cause L. The bottom left figure represents scenario 3 which is the same as scenario 1 except that K does not cause Y. The bottom right figure represents scenario 4 which is the same as scenario 2 except that X does not cause K.

Table 1 | This table displays the type-1 error rate for the test statistics using Model 1 [the Vansteelandt et al. test statistic (Vansteelandt et al., 2009)] or Model 2 (the screening test statistic) to estimate γ_1 for different allele frequencies.

Allele frequency (%)	Type-1 error rate when 1 SNP is tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.071	0.059	0.049	0.047	0.045	0.047	0.049	0.051	0.050
Scenario 1: Model 2	0.069	0.058	0.048	0.046	0.046	0.046	0.049	0.050	0.051
Scenario 2: Model 1	0.044	0.045	0.045	0.045	0.045	0.045	0.045	0.043	0.045
Scenario 2: Model 2	0.045	0.044	0.045	0.045	0.045	0.043	0.045	0.043	0.045
Scenario 3: Model 1	0.058	0.048	0.043	0.045	0.045	0.046	0.044	0.047	0.044
Scenario 3: Model 2	0.052	0.050	0.044	0.046	0.044	0.046	0.045	0.047	0.046
Scenario 4: Model 1	0.044	0.045	0.045	0.043	0.046	0.044	0.045	0.045	0.042
Scenario 4: Model 2	0.044	0.044	0.045	0.043	0.046	0.044	0.046	0.045	0.042

Table 2 | This table displays the power for the test statistics using Model 1 [the Vansteelandt et al. test statistic (Vansteelandt et al., 2009)] or Model 2 (the screening test statistic) to estimate γ_1 for different allele frequencies.

Allele frequency (%)	Power when 1 SNP is tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.264	0.363	0.448	0.504	0.576	0.629	0.669	0.692	0.706
Scenario 1: Model 2	0.241	0.361	0.444	0.508	0.581	0.633	0.671	0.696	0.710
Scenario 2: Model 1	0.180	0.302	0.406	0.492	0.564	0.610	0.649	0.667	0.686
Scenario 2: Model 2	0.180	0.302	0.408	0.491	0.563	0.610	0.646	0.666	0.685
Scenario 3: Model 1	0.265	0.365	0.449	0.504	0.581	0.632	0.669	0.696	0.712
Scenario 3: Model 2	0.246	0.361	0.451	0.510	0.586	0.634	0.671	0.699	0.716
Scenario 4: Model 1	0.175	0.304	0.408	0.499	0.558	0.607	0.647	0.671	0.681
Scenario 4: Model 2	0.174	0.303	0.407	0.498	0.557	0.605	0.648	0.672	0.682

Vansteelandt et al. (2009) has higher power and for higher allele frequencies the proposed method has higher power. However, this difference in power is negligible; the power never differs by more than 2%.

The advantage of our approach becomes clear when testing multiple SNPs. **Table 4** shows how the power to detect the causal SNP for our approach compares to Vansteelandt et al. (2009) when one SNP has a direct effect on the phenotype as simulated above in **Table 2** and 49 other SNPs are not associated with the phenotype of interest. **Table 1** shows the type-1 error rate in this scenario where the one SNP has an indirect effect on the phenotype as simulated above in **Table 1** and 49 other SNPs are not associated with the phenotype of interest or any of the other phenotypes. **Table 6** shows how the power to detect the causal SNP for our approach compares to Vansteelandt et al. (2009) when one SNP has a direct effect on the phenotype as simulated above in **Table 2** and 99 other SNPs are not associated with the phenotype of interest. **Table 5** shows the type-1 error rate in this scenario where the one SNP has an indirect effect on the phenotype as simulated above in **Table 1** and 99 other SNPs are not associated with the phenotype of interest or any of the other phenotypes.

Our approach allows for a screening step similar to the Van Steen algorithm (Van Steen et al., 2005) where the top 3 SNPs out of 50 and the top 5 SNPs out of 100 with the highest test statistic given by Equation (4) are chosen. We chose 3 SNPs out of 50 and 5 SNPs out of 100 since this is roughly 5% of the SNPs. After the top 3 or 5 SNPs are chosen based on the screening step, the test statistic described in Equation (7) is used to obtain a p-value which is compared to $\alpha/3$ and $\alpha/5$, respectively. We compare our approach with the screening step to the approach by Vansteelandt et al. (2009) with a Sidak correction. Since our approach allows for a screening step, we are better able to detect the SNP that has a direct causal effect on the target phenotype as seen in **Tables 4**, **6**.

Note that the power in **Tables 4**, **6** is lower than that in **Table 2** which is expected since multiple SNPs are tested. For more common allele frequencies, the power of using the proposed method with a screening step is more than double that of the Vansteelandt algorithm with a Sidak correction while the type-1 error rates are similar as seen in **Tables 3**, **5**. Therefore, if multiple SNPs are tested, the proposed approach has better power to detect the SNP that has a direct effect on the phenotype of interest.

Since the proposed approach is valid for testing, but may have less power when either the $X \rightarrow K$ or the $X \rightarrow Y$ link is strong, we looked at the effect of increasing the association between X and K when K influences Y ($X \rightarrow K$) and X and Y ($X \rightarrow Y$). We increased the correlation between X and K from 0.025 to 0.05 and then 0.075. We also increased the correlation between X and Y from 0.05 to 0.10 and then 0.15. The power of both statistics remained very close. At most, the power of the Vansteelandt et al. statistic (Vansteelandt et al., 2009) was 0.9% better than our approach.

Since the test statistic for the screening step given in Equation (4) is susceptible to population stratification, we examined a few scenarios where population stratification was present. We simulated half of the subjects to have allele frequency of 5, 5, 20, and 40% and the other half of the subjects to have allele frequency of 10, 45, 25, and 45%, respectively. Similar to **Tables 3**, **4**, one SNP has a direct effect on the phenotype of interest and 49 other SNPs are not associated with the phenotype of interest in

Table 3 | This table displays the significance rate when one SNP does not have a direct effect on the phenotype Y but acts as seen in Figure 2 without the arrow from X to Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Type-1 error rate when 50 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.0018	0.0008	0.0008	0.0006	0.0004	0.0006	0.0008	0.0006	0.0010
Scenario 1: Model 2	0.0014	0.0006	0.0002	0.0006	0.0012	0.0012	0.0006	0.0012	0.0006
Scenario 2: Model 1	0.0014	0.0006	0.0008	0.0012	0.0004	0.0008	0.0004	0.0008	0.0002
Scenario 2: Model 2	0.0004	0.0010	0.0012	0.0016	0.0012	0.0006	0.0010	0.0004	0.0006
Scenario 3: Model 1	0.0018	0.0006	0.0008	0.0014	0.0006	0.0010	0.0008	0.0008	0.0002
Scenario 3: Model 2	0.0014	0.0006	0.0008	0.0016	0.0012	0.0010	0.0012	0.0004	0.0006
Scenario 4: Model 1	0.0014	0.0006	0.0008	0.0012	0.0004	0.0008	0.0004	0.0008	0.0002
Scenario 4: Model 2	0.0008	0.0010	0.0013	0.0016	0.0012	0.0006	0.0010	0.0004	0.0006

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the three SNPs with the largest test statistic given by Equation (8) are tested.

Table 4 | This table displays the power when one SNP has a direct effect on the phenotype Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Power when 50 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.031	0.039	0.073	0.075	0.120	0.150	0.176	0.191	0.191
Scenario 1: Model 2	0.038	0.073	0.133	0.188	0.255	0.321	0.356	0.368	0.431
Scenario 2: Model 1	0.013	0.030	0.040	0.074	0.110	0.112	0.158	0.162	0.172
Scenario 2: Model 2	0.015	0.056	0.117	0.18	0.236	0.292	0.344	0.356	0.378
Scenario 3: Model 1	0.031	0.039	0.074	0.083	0.121	0.130	0.185	0.191	0.201
Scenario 3: Model 2	0.038	0.073	0.136	0.194	0.257	0.312	0.368	0.370	0.445
Scenario 4: Model 1	0.012	0.030	0.063	0.076	0.110	0.113	0.159	0.176	0.177
Scenario 4: Model 2	0.015	0.057	0.107	0.181	0.235	0.290	0.344	0.376	0.416

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the three SNPs with the largest test statistic given by Equation (8) are tested.

Table 5 | This table displays the significance rate when one SNP does not have a direct effect on the phenotype Y but acts as seen in Figure 2 without the arrow from X to Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Type-1 error rate when 100 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.0010	0.0006	0.0004	0.0006	0.0007	0.0006	0.0002	0.0004	0.0005
Scenario 1: Model 2	0.0008	0.0004	0.0004	0.0006	0.0006	0.0006	0.0008	0.0002	0.0006
Scenario 2: Model 1	0.0006	0.0000	0.0008	0.0000	0.0000	0.0004	0.0002	0.0006	0.0002
Scenario 2: Model 2	0.0004	0.0004	0.0008	0.0002	0.0004	0.0006	0.0010	0.0004	0.0008
Scenario 3: Model 1	0.0010	0.0010	0.0002	0.0004	0.0000	0.0004	0.0002	0.0008	0.0000
Scenario 3: Model 2	0.0008	0.0004	0.0002	0.0002	0.0002	0.0006	0.0002	0.0002	0.0004
Scenario 4: Model 1	0.0006	0.0003	0.0004	0.0006	0.0007	0.0006	0.0002	0.0004	0.0005
Scenario 4: Model 2	0.0002	0.0004	0.0004	0.0006	0.0006	0.0006	0.0008	0.0002	0.0006

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (8) are tested.

Table 6 | This table displays the power when one SNP has a direct effect on the phenotype Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Power when 100 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.014	0.028	0.049	0.048	0.084	0.109	0.111	0.147	0.142
Scenario 1: Model 2	0.021	0.056	0.099	0.136	0.196	0.262	0.277	0.332	0.351
Scenario 2: Model 1	0.004	0.018	0.040	0.055	0.076	0.099	0.098	0.116	0.123
Scenario 2: Model 2	0.014	0.042	0.088	0.145	0.178	0.246	0.249	0.284	0.332
Scenario 3: Model 1	0.018	0.028	0.038	0.049	0.087	0.094	0.112	0.128	0.139
Scenario 3: Model 2	0.023	0.057	0.099	0.137	0.198	0.229	0.283	0.315	0.368
Scenario 4: Model 1	0.006	0.018	0.040	0.041	0.076	0.086	0.098	0.116	0.123
Scenario 4: Model 2	0.011	0.042	0.088	0.126	0.178	0.209	0.249	0.284	0.332

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (8) are tested.

Table 7 | This table displays the significance level when one SNP has an indirect effect on the phenotype Y as seen in Figure 2 without the arrow from X to Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency	Type-1 error rate when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.0012	0.0014	0.0011	0.0013
Scenario 1: Model 2	0.0006	0.0006	0.0004	0.0005
Scenario 2: Model 1	0.0010	0.0006	0.0004	0.0006
Scenario 2: Model 2	0.0012	0.0013	0.0018	0.0020
Scenario 3: Model 1	0.0009	0.0002	0.0004	0.0011
Scenario 3: Model 2	0.0008	0.0012	0.0016	0.0008
Scenario 4: Model 1	0.0006	0.0014	0.0008	0.0009
Scenario 4: Model 2	0.0009	0.0006	0.0006	0.0012

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 in a screening step where the three SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

Table 8 | This table displays the power when one SNP has a direct effect on the phenotype Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency	Power when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.025	0.070	0.111	0.171
Scenario 1: Model 2	0.064	0.199	0.248	0.394
Scenario 2: Model 1	0.016	0.070	0.103	0.163
Scenario 2: Model 2	0.040	0.205	0.227	0.366
Scenario 3: Model 1	0.025	0.070	0.113	0.172
Scenario 3: Model 2	0.064	0.202	0.249	0.396
Scenario 4: Model 1	0.016	0.064	0.103	0.163
Scenario 4: Model 2	0.040	0.186	0.227	0.366

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 with a screening step where the three SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

Tables 7, **8**. Similar to **Tables 5**, **6**, one SNP has a direct effect on the phenotype of interest and 99 other SNPs are not associated with the phenotype of interest in **Tables 9**, **10**. As seen in **Tables 7**, **9**, the type-1 error rates are similar for both methods. As seen in **Tables 8**, **10**, even though there is some population stratification present, the proposed method with a screening step still performs better than the Vansteelandt algorithm, especially when the allele frequencies are more common.

DATA ANALYSIS: AN APPLICATION TO THE FRAMINGHAM STUDY

We evaluated the practical relevance of the proposed adjustment principle by an application to the Framingham Heart Study with 1400 probands (Herbert et al., 2006). For the target phenotype, we selected the lung-function measurement FEV1. For the secondary phenotype K, we selected height. Gender, and age represent L, the collection of common risk factors between FEV1 and height. For rs2415815 a SNP associated with both height and FEV1, the test statistic equals 0.044 with corresponding p-value equal 0.83. As a result, we fail to reject the null hypothesis and conclude that there is no evidence that the SNP acts directly on FEV1 other than via body height.

DISCUSSION

Our proposed FBAT assesses the direct genetic effect of a marker locus on the phenotype of interest, other than through another correlated phenotype. The adjustment is based on the conditional mean model approach and can be incorporated into the FBAT-approach in a straightforward fashion. The power of the approach is assessed by simulation studies and shown to be similar to the Vansteelandt et al. method when only one SNP is being tested and superior when multiple SNPs are being tested (Vansteelandt et al., 2009). Unlike the Vansteelandt et al. method, this method

Table 9 | This table displays the significance level when one SNP has an indirect effect on the phenotype Y as seen in Figure 2 without the arrow from X to Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency	Type-1 error rate when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.0011	0.0005	0.0007	0.0003
Scenario 1: Model 2	0.0009	0.0006	0.0008	0.0003
Scenario 2: Model 1	0.0004	0.0015	0.0009	0.0005
Scenario 2: Model 2	0.0003	0.0011	0.0012	0.0005
Scenario 3: Model 1	0.0004	0.0010	0.0008	0.0004
Scenario 3: Model 2	0.0006	0.0009	0.0010	0.0006
Scenario 4: Model 1	0.0008	0.0013	0.0007	0.0004
Scenario 4: Model 2	0.0010	0.0008	0.0011	0.0006

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

Table 10 | This table displays the power when one SNP has a direct effect on the phenotype Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency	Power when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.022	0.050	0.073	0.157
Scenario 1: Model 2	0.044	0.141	0.170	0.324
Scenario 2: Model 1	0.014	0.046	0.071	0.148
Scenario 2: Model 2	0.036	0.137	0.161	0.298
Scenario 3: Model 1	0.022	0.050	0.076	0.159
Scenario 3: Model 2	0.045	0.143	0.174	0.326
Scenario 4: Model 1	0.014	0.046	0.071	0.148
Scenario 4: Model 2	0.036	0.137	0.161	0.298

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

uses a screening step and has the unique advantage in situations in which a large number of SNPs are tested for a direct effect on the phenotype of interest. Since the number of tests will be much smaller than the total number of SNPs, this will lead to substantial reduction in the adjustment for multiple-comparisons and will result in improved overall statistical power. In this process, the screening step works under the assumption of no population admixture, but the final analysis of the selected SNPs is robust against it.

While we considered several causal scenarios, if the causal relationships assumed in the DAGs are not true this could cause problems for the proposed method. For example, a causal arrow $K \leftarrow Y$ or $L \rightarrow Y$ could introduce spurious association for this method. Therefore, one needs to makes sure that the assumptions of the DAG are met before using the proposed approach. While the simulations considered 50 and 100 SNPs, a realistic application could involve thousands of GWAS SNPs. This leads to extreme multiple test corrections and may lead to very different behavior than the behavior observed in the simulation studies (Morris and Elston, 2011). Furthermore, if phenotypes of the founders are known, the proposed method could perform poorly compared to population-based approaches.

For the screening step in the Simulations section, we chose 3 out of 50 and 5 out of 100 SNPs since this is roughly 5% of the tested SNPs. Another number of SNPs could be chosen for the screening step. Although, if the majority of SNPs are chosen in the screening step (i.e., 40 out of 50 SNPs), this increases the number of multiple comparisons and can decrease power. If too few SNPs are chosen in the screening step (i.e., 1 out of 50 SNPs), this decreases the number of multiple comparisons, but one may fail to detect the causal SNP since too few SNPs were chosen. Care needs to be given to the number of SNPs chosen in the screening step (Van Steen et al., 2005). One cannot simply choose different numbers of SNPs for the screening step until significant results are found since this will inflate the type-1 error rate (Van Steen et al., 2005).

ACKNOWLEDGMENTS

This work was funded by NIH/NHLBI U01 HL089856 Edwin K. Silverman, PI. COPDGene is supported by NHLBI Grant Nos U01HL089897 and U01Hl089856. This work was also funded by FWO grant G.0111.12.

REFERENCES

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. *Nat. Genet.* 40, 616–622. doi: 10.1038/ng.109

Berzuini, C., Vansteelandt, S., Foco, L., Pastorino, R., and Bernardinelli, L. (2012). Direct genetic effects and their estimation from matched case-control data. *Genet. Epidemiol.* 36, 652–662. doi: 10.1002/gepi.21660

Cole, S. R., and Hernan, M. A. (2002). Fallibility in estimating direct effects. *Int. J. Epidemiol.* 31, 163–165. doi: 10.1093/ije/31.1.163

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., et al. (2006). A common genetic variant is associated with adult and childhood obesity. *Science* 312, 279–283. doi: 10.1126/science.1124779

Lange, C., DeMeo, D., Silverman, E. K., Weiss, S. T., and Laird, N. M. (2003a). Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am. J. Hum. Genet.* 73, 801–811. doi: 10.1086/378591

Lange, C., Lyon, H., DeMeo, D., Raby, B., Silverman, E. K., and Weiss, S. T. (2003b). A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. *Hum. Hered.* 56, 10–17. doi: 10.1159/000073728

Morris, N., and Elston, R. (2011). A note on comparing the power of test statistics at low significance levels. *Am. Stat.* 65, 164. doi: 10.1198/tast.2011.10117

Murphy, A., Weiss, S. T., and Lange, C. (2008). Screening and replication using the same data set: testing strategies for family-based studies in which all probands are affected. *PLoS Genet.* 4:e1000197. doi: 10.1371/journal.pgen.1000197

Rabinowitz, D., and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* 50, 211–223. doi: 10.1159/000022918

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452, 638–642. doi: 10.1038/nature06846

VanderWeele, T. J., Asomaning, K., Tchetgen, E. J., Han, Y., Spitz, M. R., Shete, S., et al. (2012). Genetic variants on 15q25.1, smoking and lung cancer: an assessment of mediation and interaction. *Am. J. Epidemiol.* 175, 1013–1020. doi: 10.1093/aje/kwr467

Vansteelandt, S. (2010). Estimating direct effects in cohort and case-control studies. *Epidemiology* 21, 278. doi: 10.1097/EDE.0b013e3181cd72aa

Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., et al. (2009). On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. *Genet. Epidemiol.* 33, 394–405. doi: 10.1002/gepi.20393

Vansteelandt, S., and Lange, C. (2012). Causation and causal inference for genetic effects. *Hum. Genet.* 131,1665–1676. doi: 10.1007/s00439-012-1208-9

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B., Lyon, H., Demeo, D. L., et al. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37, 683–691. doi: 10.1038/ng1582

Won, S., Bertram, L., Becker, D., Tanzi, R. E., and Lange, C. (2009). Maximizing the power of genome-wide association studies: a novel class ofpowerful family-based association tests. *Stat. Biosci.* 1, 125–143. doi: 10.1007/s12561-009-9016-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 June 2013; accepted: 25 October 2013; published online: 21 November 2013.

Citation: Lutz SM, Vansteelandt S and Lange C (2013) Testing for direct genetic effects using a screening step in family-based association studies. Front. Genet. ***4****:243. doi: 10.3389/fgene.2013.00243*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

APPENDIX

The following proof shows that the test statistics in the first and second screening steps are uncorrelated under the null hypothesis. As discussed in the introduction and methods sections, $\tilde{Y} = Y - \bar{y} - \gamma_1 K - \bar{k}$ is the adjusted phenotype for the effect that the phenotype K has on the target phenotype Y. For ease of notation, we will use $\tilde{Y} = Y - \gamma_1(K)$ for this proof. Suppose that the null hypothesis is true that X has no effect on Y other than through K. Let

$$E(Y|X, K, U) = E(Y|K, U) = \Phi\{w(U) + \gamma_1 K\} \tag{9}$$

where Φ equals the identity link or exponential link and $w(U)$ is an arbitrary function. Without loss of generality, for the following proof, let Φ equal the identity link. This model does not involve X because we are working under the null hypothesis of no direct effect. Furthermore, the parameter γ_1 in this model is the same as in model

$$E(Y|X, K, L, S) = w^*(X, L, S) + \gamma_1 K \tag{10}$$

for some function $w^*(X, L, S)$ of (X, L, S), which can be seen by inferring this model from model (9) upon noting that $Y \perp\!\!\!\perp (L, S)|X, K, U$ and $U \perp\!\!\!\perp K|L, X, S$. Using model (9) and model (10) and noting that $Y \perp\!\!\!\perp S|X, K, U$ and $X \perp\!\!\!\perp U|S$, then

$$\begin{aligned} E[\tilde{Y}(X - E[X|S])] &= E[(Y - \gamma_1 K)(X - E[X|S])] \\ &= E[w(U)(E[X|S, U] - E[X|S])] = 0 \end{aligned} \tag{11}$$

As a result of Equation (11) and model (9), the $\mathrm{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S])$ can be written as follows

$$\begin{aligned} &\mathrm{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S]) \\ &\quad = E[(Y - \gamma_1 K)^2 E(X|S)(X - E[X|S])] \\ &\quad = E[(Y - E(Y|X, K, U) + w(U))^2 E(X|S)(X - E[X|S])] \\ &\quad = \mathrm{Part}_1 + \mathrm{Part}_2 + \mathrm{Part}_3 \end{aligned}$$

where

$$\begin{aligned} \mathrm{Part}_1 &= E[\left(w(U)^2\right) E[X|S](X - E[X|S])] \\ \mathrm{Part}_2 &= E[(2(w(U)(Y - E[Y|X, K, U])) E[X|S](X - E[X|S])] \\ \mathrm{Part}_3 &= E[\left((Y - E[Y|X, K, U])^2\right) E[X|S](X - E[X|S])] \end{aligned} \tag{12}$$

We will show that the $\mathrm{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S]) = 0$ by showing that each part of the above equation equals zero.

$$\begin{aligned} \mathrm{Part}_1 &= E[w(U)^2 E[X|S](X - E[X|S])] \\ &= E[w(U)^2 E[X|S](E[X|S, U] - E[X|S])] = 0 \end{aligned} \tag{13}$$

because $X \perp\!\!\!\perp U|S$.

$$\mathrm{Part}_2 = E[2w(U)(Y - E[Y|X, K, U])E[X|S](X - E[X|S])] = 0 \tag{14}$$

because $Y \perp\!\!\!\perp S|X, K, U$.

$$\begin{aligned} \mathrm{Part}_3 &= E[(Y - E[Y|X, K, U])^2 E[X|S](X - E[X|S])] \\ &= E[E[X|S](X - E[X|S])\,Var[Y|K, U]] \end{aligned} \tag{15}$$

because $Y \perp\!\!\!\perp S|X, K, U$ and $Y \perp\!\!\!\perp X|K, U$.

Assuming that $\mathrm{Var}(Y|K, U)$ is constant, as we do throughout, it is immediate that the term Part_3 is zero. As a result, this shows that $\mathrm{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S]) = 0$.