RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Robust and Powerful Affected Sibpair Test for Rare Variant Association

Keng-Han Lin[1,2] and Sebastian Zöllner[1,2,3]*

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States of America; [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America; [3]Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, United States of America

Received 5 November 2014; Revised 25 March 2015; accepted revised manuscript 1 April 2015.
Published online 13 May 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21903

ABSTRACT: Advances in DNA sequencing technology facilitate investigating the impact of rare variants on complex diseases. However, using a conventional case-control design, large samples are needed to capture enough rare variants to achieve sufficient power for testing the association between suspected loci and complex diseases. In such large samples, population stratification may easily cause spurious signals. One approach to overcome stratification is to use a family-based design. For rare variants, this strategy is especially appropriate, as power can be increased considerably by analyzing cases with affected relatives. We propose a novel framework for association testing in affected sibpairs by comparing the allele count of rare variants on chromosome regions shared identical by descent to the allele count of rare variants on nonshared chromosome regions, referred to as test for rare variant association with family-based internal control (TRAFIC). This design is generally robust to population stratification as cases and controls are matched within each sibpair. We evaluate the power analytically using general model for effect size of rare variants. For the same number of genotyped people, TRAFIC shows superior power over the conventional case-control study for variants with summed risk allele frequency $f < 0.05$; this power advantage is even more substantial when considering allelic heterogeneity. For complex models of gene-gene interaction, this power advantage depends on the direction of interaction and overall heritability. In sum, we introduce a new method for analyzing rare variants in affected sibpairs that is robust to population stratification, and provide freely available software.
Genet Epidemiol 39:325–333, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: rare variants; dichotomous traits; family studies; association test; sequencing

Introduction

Rare variants with large relative risk are hypothesized to explain some of the missing heritability of complex diseases [Manolio et al., 2009]. Several studies have identified rare variants underlying rare Mendelian diseases using next-generation sequencing technology [Ng et al., 2010; Ng et al., 2009]. However, the conventional case-control design has low statistical power to detect the association between rare variants and complex diseases [Cooper and Shendure, 2011; Li and Leal, 2008]. To overcome the low power of single-marker test on rare variants, researchers have proposed to combine variants in a gene or genomic region to test for association [Li and Leal, 2008; Price et al., 2010; Zawistowski et al., 2014]. However, such gene-based tests in population samples may still need >10,000 individuals to identify the signal from rare variants [Nelson et al., 2012]; sequencing such large samples is still very expensive. Moreover, large samples are typically more heterogeneous in origin, increasing the risk of population stratification [Price et al., 2006]. In such large samples, even subtle stratification causes substantially increased false-positive rate in rare variant tests [Zawistowski et al., 2014]. Although methods to control for population stratification, such as principal components and genomic control [Devlin and Roeder, 1999; Price et al., 2006], have been successfully applied for common variants, it is unclear whether such methods are appropriate for rare variant tests [Liu, Nicolae and Chen, 2013; Mathieson and McVean, 2012].

As family members are naturally matched for genetic background, several recent gene-based methods for testing the association between rare variants and the phenotype adapt family data to control for population stratification [De et al., 2013; Guo and Shugart, 2012]. In addition, the allele frequency of rare risk variants in cases can be substantially increased by collecting cases with affected relatives [Fingerlin et al., 2004; Peng et al., 2010; Zöllner, 2012]. Although collecting families with multiple affected members is challenging, family-based studies of rare variants can leverage existing large collections of families that were originally generated for linkage analysis [Guan et al., 2012; Howson et al., 2009; Rao et al., 2003]; for example, International Type 2 Diabetes Linkage Analysis Consortium contains >4,000 affected sibpairs [Guan et al., 2012].

Methods have been proposed to extend the current collapsing tests to rare variants in family data. Guo and Shugart [2012] and De et al. [2013], extended the family-based

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Sebastian Zöllner, Department of Biostatistics, University of Michigan, 1420 Washington Heights, SPH II, Ann Arbor, MI 48109, USA. E-mail: szoellne@umich.edu

association test (FBAT) [Laird and Lange, 2006] to rare variants in the style of a collapsing test. Schifano et al. 2012 and Chen et al. 2013 used linear mixed models to extend the SNP-set kernel association test (SKAT) [Wu et al., 2011] to families. Shugart et al. [2012] and Fang et al. [2012] proposed to estimate the relatedness between samples and adjust the test statistics for rare variant association accordingly. However, none of the existing methods directly leverage the benefit of studying families where the same rare variant is observed multiple times. By using such information, we can increase power to detect the association between rare variants and the phenotype.

Here, we propose a powerful framework for testing rare variant associations using affected sibpairs. We create a matched design by comparing the allele count of rare variants on shared identity by descent (IBD) chromosome regions to the allele count on nonshared identity by descent chromosome regions across affected sibpairs in a region of interest. Sharing status of chromosome regions can be easily estimated using high-density genotype data [Keith et al., 2008], and sharing status of alleles can be inferred conditional on the known chromosome region sharing status. Intuitively, we consider shared chromosome regions as "case" chromosome regions and nonshared chromosome regions as "control" chromosome regions. Under the null hypothesis of no association, the probability of a shared chromosome region carrying an allele is identical to the probability of a nonshared chromosome region carrying an allele. Under the alternative that an allele increases/decreases the disease risk, the probability of a shared chromosome region carrying that allele is higher/lower than the probability of a nonshared chromosome region carrying that allele.

We evaluate this design by calculating the analytical power for a collapsing gene-based test [Li and Leal, 2008], assuming a general model of rare risk alleles that is specified by the summed allele frequency of all rare risk variants in the gene and the mean and variance of their effect size [Zöllner, 2012]. We show that given the same number of sequenced individuals, the power of the proposed affected sibpair test for rare variant association with family-based internal control (TRAFIC) is higher than the conventional case-control design for rare risk variants (summed risk allele frequency < 0.05). Considering allelic heterogeneity, where risk variants have different effect sizes, TRAFIC doubles the power of a case-control study in many realistic parameter values. We also evaluate the power of the proposed method under various gene-gene interaction models and find that power depends on the type of interaction and the overall heritability of the disease. Using simulations, we also show that the proposed TRAFIC is generally robust to population stratification.

Materials and Methods

TRAFIC

We consider a set of affected sibpairs with known number of chromosome regions shared identical by descent (IBD).

Table 1. Identification of variant IBD status conditional on chromosome region IBD status

	0 IBD chromosome region	1 IBD chromosome region	2 IBD chromosome regions
Both siblings are homozygous minor allele	Four nonshared alleles	One shared and 2 nonshared alleles	Two shared alleles
One homozygous minor allele and one heterozygote	Three nonshared alleles	One shared and 1 nonshared alleles	N/A
Both siblings are heterozygous	Two nonshared alleles	Ambiguous configuration	One shared allele

Assuming chromosome region IBD status is known, the number of shared and nonshared alleles can be inferred for all but one configuration of genotypes (shaded cell).

At a locus of interest (e.g., a gene), we compare the count of alleles of rare variants on chromosome regions shared IBD between the siblings to the count of alleles of rare variants on chromosome regions not shared IBD (non-IBD chromosome regions) across sibpairs. Let, p^{IBD} be the frequency of IBD chromosome region carrying at least one allele and p^{NonIBD} be the frequency of non-IBD chromosome regions carrying at least one allele. Alleles without effect on disease risk are equally likely to occur on any chromosome region regardless of IBD status. Thus, the null hypothesis under no association is $H_0 : p^{IBD} = p^{NonIBD}$. Variants that are associated with the phenotype (protective or causative) would differ in frequency between IBD and non-IBD chromosome regions. Hence, we can test for departure from the null hypothesis either in a collapsing framework by considering the alternative $H_a : p^{IBD} \neq p^{NonIBD}$ or in a dispersion framework where this alternative is considered for each variant and the combined test statistic aggregates the evidence across all variants.

In a sibpair with known IBD status, identifying whether an allele of a variant is located on an IBD or a non-IBD chromosome region is straightforward for most genotypes as shown in Table 1; for example, when a sibpair does not share the chromosome region (0 IBD chromosome region), all observed alleles for that variant in two siblings are nonshared; for a sibpair who shares 1 IBD chromosome region, the alleles of a homozygous sibling must be one shared and one nonshared. Only when the sibpair shares one IBD chromosome region and the genotypes are heterozygous in both individuals, the IBD status of the allele is ambiguous (shaded in Table 1): this configuration could be either the result of a single rare allele located on the IBD chromosome region or two copies of the rare allele inherited separately on the non-IBD chromosome regions (as illustrated in supplementary Appendix Fig. S1). To resolve this ambiguous configuration, we implement an imputation algorithm and use simulations to show the false-positive rate is controlled (see supplementary Appendix 1 for details).

Evaluating TRAFIC

The analytical power of the proposed TRAFIC based on a collapsing gene-based test depends on the difference between

the expected allele count on shared IBD chromosome regions and the expected allele count on nonshared IBD chromosome regions. To calculate these expectations, we assume that all rare variants evaluated in a locus occur on different haplotypes. Let f be the sum of population allele frequencies of all risk variants (summed risk allele frequency). For each sibpair, we count the number of alleles $H_S \in \{0, 1, 2\}$ on the shared chromosome regions and the number of alleles $H_{NS} \in \{0, 1, 2, 3, 4\}$ on nonshared chromosome regions. Let AA_R be an affected sibpair and $P(H_S, H_{NS}|AA_R, S)$ be the probability of H_S, H_{NS} conditional on the number of shared IBD chromosome regions $S \in \{0, 1, 2\}$. Using Bayes' rule, we can write this conditional probability as

$$P(H_S, H_{NS}|AA_R, S) = P(AA_R|H_S, H_{NS})\, P(H_S, H_{NS}|S) \times P(S)\, \frac{1}{P(AA_R, S)},$$

where $P(AA_R|H_S, H_{NS})$ depends on the underlying genetic and effect size model (see supplementary Appendix 2 for derivations). Based on previous work [Zöllner, 2012], we model the effect size (relative risk) of each risk haplotype as a random variable with the first two moments μ and σ^2. Then, $P(H_S, H_{NS}|AA_R, S)$ is fully determined by the parameters μ, σ^2, and f (see supplementary Appendix 2). We calculate the power for TRAFIC based on $P(H_S, H_{NS}|AA_R, S)$ for a range of relative risk parameter μ and σ^2, and under different f assuming a simple collapsing method [Li and Leal, 2008] to test the association between rare variants and the dichotomous phenotype (supplementary Appendix 3). To maintain an overall false-positive rate of 0.05 after testing 20,000 genes in the genome, we set the false-positive rate to 2.5×10^{-6}. We compare our proposed TRAFIC with two other designs: (1) the conventional case-control study comparing a sample of cases to unaffected controls. (2) A selected cases design comparing cases that are ascertained to have an affected sibling to unaffected controls [Fingerlin et al., 2004, Zöllner, 2012]. All designs retain the nominal false-positive rate under the null (supplementary Appendix Table S1).

Simulation Setup for TRAFIC

To validate the derived analytical results, we simulate sibpair samples and apply our proposed TRAFIC. We first generate four independent parental haplotypes, each carrying a risk allele with probability f. Without considering recombination, we then generate two descendants, each randomly inheriting one chromosome region from each parent. Following Risch [1990], we define the contribution to prevalence K at the locus of interest as K_L and the contribution of the remaining genome as K_G. The prevalence among subjects with an affected relative with relation status R is K_R; the contribution to K_R at the locus of interest and the remaining genome are then K_{LR} and K_{GR}, respectively. We adjust $K_G K_{GR}$ under the multiplicative model to maintain both K and the sibling relative risk (SRR).

$$SRR = \frac{K_L K_{LR} K_G K_{GR}}{K \times K}.$$

Here $K_L K_{LR}$ depends on $P(AA_R|H_S, H_{NS})$ (more details in supplementary Appendix 2). The relative risk of the risk allele follows a gamma distribution with specified μ and σ^2. Thus, the probability of having both siblings in the family affected is $K_L K_{LR} K_G K_{GR}$ and is set to 1 if the simulated probability exceeds 1. We generate datasets of 1,000 affected sibpairs in each replicate. To evaluate the performance of our multiple imputation algorithm, we generate sibpairs assuming the sharing status is known. Then, we mask the true location for the double-heterozygote sibpairs who share one IBD chromosome region and apply our multiple imputation algorithm.

Population Stratification

Using the simulation design described above, we evaluate the impact of population stratification. We simulate two populations with summed risk allele frequency of 0.01 and 0.05, respectively, and assign a ratio of prevalence π between two populations. Assuming two populations have the same sibling relative risk, the ratio of frequencies of affected sibpairs between the two populations is then π^2. Assuming that both populations contribute equally, we generate case-control samples by sampling 1,000 cases, a proportion of $\pi/(1+\pi)$ from population 1 and $1/(1+\pi)$ from population 2. We also sample 1,000 controls with equal contribution from each population. To generate a stratified sample for TRAFIC, we generate a sample of 1,000 affected sibpairs with a proportion of $\pi^2/(1+\pi^2)$ from population 1 and a proportion of $1/(1+\pi^2)$ from population 2. We assume unknown sharing status for double-heterozygote sibpairs who share one IBD chromosome region and impute the sharing status through multiple imputation. To generate cases for the selected cases design, we sample affected sibpairs with a proportion of $\pi^2/(1+\pi^2)$ from population 1 and $1/(1+\pi^2)$ from population 2; controls are sampled evenly from both populations. We generate 1,000 datasets for each value of π and estimate the false-positive rate.

Gene-Gene Interaction

Interaction between the locus of interest and the remaining genome can influence the power of association tests in family samples [Risch, 2001; Zöllner, 2012]. We model gene-genome interaction as two loci, L and G. L is the locus of interest, while G represents genetic effects in the remainder of the genome. We define the joint effect as

$$P(A|h_m, h_n, g_s, g_t) \propto \beta_L^{h_m+h_n}\, \beta_G^{g_s+g_t}\, \gamma^{(h_m+h_n)(g_s+g_t)}.$$

where h_m and h_n represent the indicator of a risk allele at locus L; let g_s and g_t represent the indicator of a risk allele at locus G. In the absence of risk alleles at G, all risk alleles at locus L have the same relative risk β_L. Moreover, we describe the extent of interaction in this model by the parameter γ as the relative risk when risk alleles are present at both loci L and G, where $\gamma = 1$ indicates no interaction, $\gamma < 1$ indicates antagonistic interaction, and $\gamma > 1$ indicates synergistic interaction.

Under this model, the marginal relative risk at locus L is

$$\frac{P(A\,|h_m=1)}{P(A\,|h_m=0)} = \frac{\beta_L \sum_{h_n} \beta_L^{h_n} \rho(h_n) \sum_{gs} \sum_{gt} (\beta_G \gamma)^{g_s+g_t} \gamma^{h_n(g_s+g_t)} p(g_s) p(g_t)}{\sum_{h_n} \beta_L^{h_n} \rho(h_n) \sum_{g_s} \sum_{g_t} \beta_G^{g_s+gt} \gamma^{h_n(g_s+g_t)} p(g_s) p(g_t)}.$$

The marginal relative risk at locus G is expressed in a similar fashion. To explore the effect of gene-gene interaction, given the sibling relative risk, we vary γ while adjusting β_L and β_G to keep the marginal relative risks constant (see supplementary Appendix 4). This maintains a constant power for the conventional case-control study. We then calculate $P(H_S, H_{NS}|AA_R)$ at locus L and evaluate the power of TRAFIC for different values of γ.

An Example to Illustrate TRAFIC

To illustrate how to apply TRAFIC, we simulate 1,000 sibpairs assuming the number of shared IBD chromosome region is known. We simulate sequence data by using coalescent model based simulator COSI [Schaffner et al., 2005] to generate a population of ten thousand 1 kb haplotypes. From the 50 variants in the region, we randomly pick 10 variants with minor allele frequency (MAF) < 0.05 and assign each variant the relative risk as a function of MAF, $-\log_{10}(\text{MAF})$ [Wu et al., 2011]. In this setting, a variant with MAF = 0.05 has relative risk of 1.33 and a singleton has relative risk of 4. We thus generated a population with $f = 0.025$, $\mu = 2.52$, and $\sigma^2 = 0.62$. We then generate 1,000 affected sibpairs and apply TRAFIC to that dataset.

The simulated data contain 254, 509, and 237 sibpairs who share 0, 1, and 2 chromosome regions, respectively; these equal to 983 shared chromosome regions and 2,034 nonshared chromosome regions. Excluding 42 sibpairs who shared one chromosome region with ambiguous double-heterozygote genotypes, there are 51 shared and 67 nonshared chromosome regions carrying at least one allele (carrier). Using imputation to resolve the IBD status of allele from 42 sibpairs with ambiguous double-heterozygote genotypes, the mean count of carrier chromosome regions is 91.7 on shared chromosome regions and 67.6 on nonshared chromosome regions. Using a χ^2 test, we reject the null hypothesis that IBD and non-IBD chromosome regions are equally likely to carry at least one allele ($P = 5.63 \times 10^{-11}$) indicating the presence of risk variants at this locus.

Results

We proposed a new gene-based method for analyzing affected sibpairs by comparing the risk alleles on shared IBD chromosome regions with the risk alleles on nonshared IBD chromosome regions. We evaluated the proposed TRAFIC design assuming a collapsing gene-based test by modeling allelic heterogeneity at the locus of interest based on a summed allele frequency of all risk variants f and a distribution of effect sizes with mean μ and variance σ^2. For comparison, we also evaluated the conventional cases-control design (conventional) and a case-control design in which the cases are selected conditional on having an affected sibling (selected cases) under the same genetic model. For all three designs, we assumed equal number of sequenced or genotyped individuals. To use consistent language, we referred to shared IBD chromosome regions in TRAFIC as cases and to nonshared IBD chromosome regions as controls.

First, we compared the expected summed minor allele frequency (sMAF) in cases and controls with and without allelic heterogeneity to illustrate how TRAFIC behaved relative to the conventional and selected cases designs. We then calculated the analytical power of three designs. We also evaluated robustness to population stratification. Finally, we calculated the analytical power of TRAFIC while considering different directions of gene-gene interaction.

Frequency Distribution of Risk Variants

To quantify the enrichment of risk variants in TRAFIC, we calculated the expected sMAF of risk variants in cases and controls of TRAFIC for a range of genetic models (see supplementary Appendix 3 for details). Initially, we modeled a locus with constant genetic risk μ between 1 and 5 for all variants ($\sigma^2 = 0$) (Fig. 1) and a disease prevalence of 0.01. In TRAFIC (Fig. 1A), sMAF increased rapidly in cases (shared IBD chromosome regions) and also increased roughly linearly with μ in controls (nonshared IBD chromosome regions). In the conventional design (Fig. 1B), sMAF increased in cases almost linearly with relative risk, only slightly faster than the sMAF in controls of TRAFIC. In the selected cases design (Fig. 1C), sMAF in cases with affected siblings increased faster than cases in the conventional case-control design but slower than sMAF in cases of TRAFIC. Both in the conventional design and selected cases design, sMAF in controls decreased slightly as μ increased, especially for more common variants ($f = 0.20$). As a result, TRAFIC generated a larger difference in sMAF between cases and controls than the conventional case-control design in models with $f = 0.01$ and 0.05. This advantage of TRAFIC reduced with increasing f. For $\mu = 2$, the difference in sMAF of TRAFIC compared to the conventional design was 190% (0.019–0.010) at $f = 0.01$ and reduced to 123% (0.166–0.135) at $f = 0.20$. For a higher disease prevalence of 0.20, the sMAF in controls decreased more rapidly as μ increased and the difference between cases and controls grew further in the conventional case-control and selected cases design (supplementary Appendix Fig. S2).

To evaluate scenarios where genetic effect differs between risk variants, we considered a distribution of relative risks with $\sigma^2 > 0$ while maintaining $\mu = 1.5$ (Fig. 2); for $f = 0.01$, a value $\sigma^2 = 0.1$ represents, for example, a scenario of 20 tested variants with equal frequencies where 6 of the tested variants are nonfunctional (relative risk = 1) and 14 of the tested variants have a relative risk of 1.71. A value $\sigma^2 = 0.2$ would, for example, represent 9 nonfunctional variants and

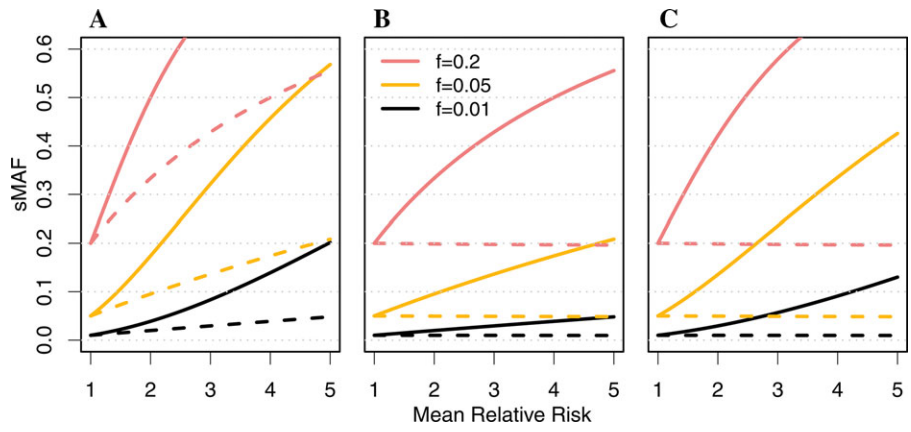

Figure 1. Summed minor allele frequency (sMAF) of risk variants in cases (solid lines) and controls (broken lines) under different study designs. We show sMAF as a function of mean relative risk of risk variants for (A) TRAFIC, (B) the conventional case-control design, and (C) the selected cases design for summed allele frequencies (*f*) of 0.01, 0.05, and 0.2 and fixed variance of relative risk $\sigma^2 = 0$.

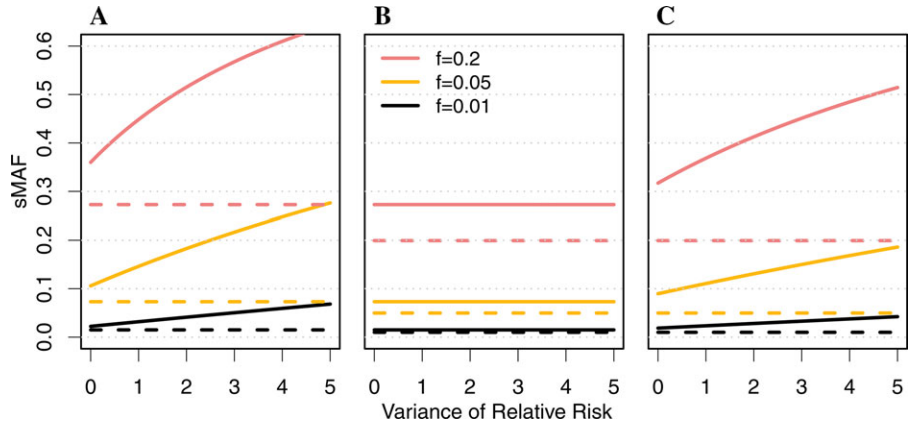

Figure 2. Summed minor allele frequency (sMAF) of risk variants in cases (solid lines) and controls (broken lines) under different study designs. We show sMAF as a function of variance of relative risk between risk variants for (A) TRAFIC, (B) the conventional case-control design, and (C) the selected cases design for summed allele frequencies (*f*) of 0.01, 0.05, and 0.2 and fixed mean relative risk $\mu = 1.5$.

11 variants with relative risk 1.91. Increasing σ^2 did not affect sMAF in cases or controls in the conventional design, as in this design sMAFs only depended on μ (Fig. 2B). In TRAFIC, sMAF in cases increased with σ^2 while the sMAF in controls remained constant. Similarly, in the selected cases design, sMAF in cases increased with σ^2, albeit more slowly than for TRAFIC (Fig. 2A and C). Even if the average effect of risk variants is 1 ($\mu = 1$), the difference in sMAF between cases and control increases with growing σ^2 for TRAFIC and for the selected cases design (supplementary Appendix Fig. S3).

Power Analysis

Based on the differences in expected sMAF, we calculated the analytical power for three study designs for the same number of individuals ($n = 2{,}000$): (1) 1,000 affected sib-pairs using TRAFIC, (2) 1,000 cases and 1,000 controls in the conventional cases-control design, and (3) 1,000 cases with affected siblings and 1,000 controls in the selected cases design. Thus, we generated 4,000 independent observations for the conventional and the selected design, and ~3,000 independent observations (~1,000 cases and ~2,000 controls) for TRAFIC. We also determined power empirically using simulations and observed no difference between empirical power and analytical power (supplementary Appendix Fig. S4).

Assuming all risk variants had the same relative risk between 1 and 5 ($\sigma^2 = 0$), the selected cases design was uniformly most powerful (Fig. 3A) while the power ranking of TRAFIC and the conventional design depended on *f*. For rarer risk variants ($f < 0.05$), TRAFIC had substantially higher power than the conventional design across all relative risks analyzed. For example, for $f = 0.01$ and $\mu = 2.5$, the power of the conventional design was 0.131 compared to 0.532 for TRAFIC. With increasing *f* or increasing prevalence, the power difference between TRAFIC and the conventional design reduced. For sets of risk variants with $f > 0.05$, the

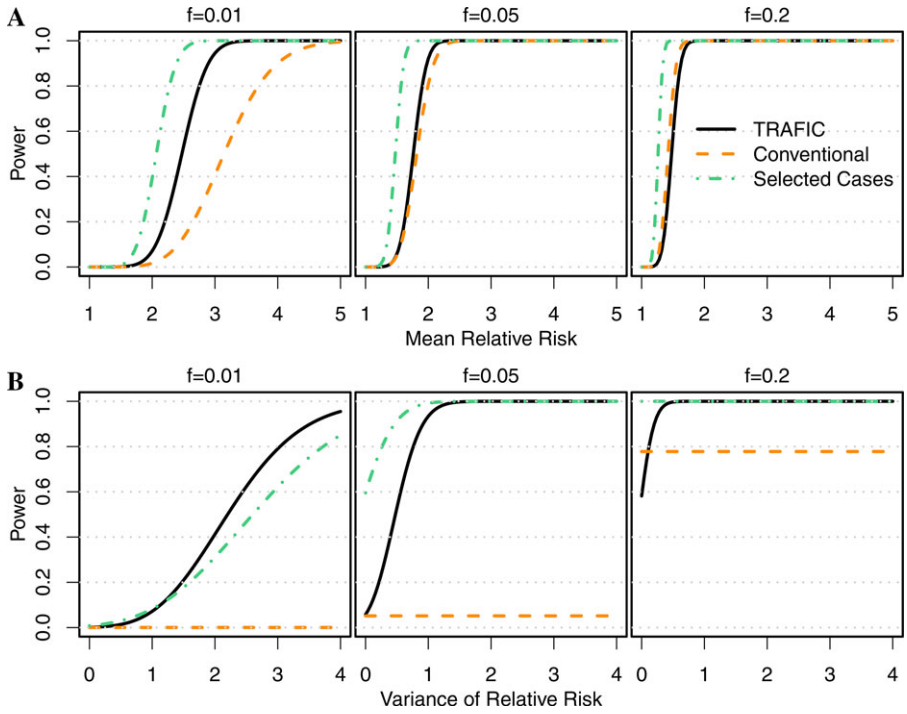

Figure 3. The analytical power curve for TRAFIC, conventional case-control, and selected cases design for different summed allele frequencies (f). Row (A) displays the power as a function of mean relative risk evaluated at variance of relative risk $\sigma^2 = 0$. Row (B) shows the power as a function of variance of relative risk evaluated at mean relative risk $\mu = 1.5$. Results are shown for 2,000 individuals (1,000 sibpairs or 1,000 cases and 1,000 controls) at a significance level 2.5×10^{-6}.

power of the conventional design was larger than the power of TRAFIC. The ranking of TRAFIC with the conventional design depended on the prevalence of the disease, for prevalence 0.20, the conventional design was already more powerful than TRAFIC for $f > 0.01$ (supplementary Appendix Fig. S5).

In models with allelic heterogeneity ($\sigma^2 > 0$), power of TRAFIC increased with rising σ^2 while the power of the conventional design was independent of σ^2 and only depended on f (Fig. 3B). For $f = 0.01$ and 0.05 at $\mu = 1.5$, the power of TRAFIC was uniformly greater than the power of the conventional design. For $f = 0.2$, TRAFIC was more powerful than the conventional design for $\sigma^2 > 0.1$. Even for high-prevalence diseases, TRAFIC is more powerful than the conventional design at modest levels of heterogeneity (supplementary Appendix Fig. S5). Moreover, the selected cases design was no longer uniformly most powerful in the presence of moderate allelic heterogeneity. For example, when $f = 0.01$ and $\sigma^2 = 2$, TRAFIC outperformed the selected cases design (with power of 0.412 and 0.306, respectively). For a model with no mean effect ($\mu = 1$), TRAFIC was uniformly most powerful regardless of f (results not shown).

Population Stratification

We modeled the level of population stratification by the parameter π that represents the ratio of prevalence between two populations (see Materials and Methods). In the absence of true risk variants ($\mu = 1$, $\sigma^2 = 0$), the conventional case-control design and the selected cases design only achieved the nominal false-positive rate at $\pi = 1$ where equal proportion of cases and controls are sampled from the two populations. Both designs showed substantially increased false-positive rate when moving away from $\pi = 1$. Especially, the selected cases design showed a high false-positive rate for moderate levels of stratification. For $\pi = 1.22$, the false-positive rate was 0.064 and 0.107 for the conventional case-control and selected cases designs; the inflation increased to 0.725 and 0.973 when $\pi = 4.06$. TRAFIC maintained the false-positive rate at the nominal level of 0.05 across the range of π (Fig. 4) as long as we assumed either no linkage signal or a linkage signal of the same strength in the two populations. If we model a strong linkage signal in only one of the populations, we observe a slightly increased false-positive rate in TRAFIC (supplementary Appendix 5).

Gene-Gene Interaction

We summarized the effect of the gene-gene interaction in a two-locus model by the parameter γ (see Materials and Methods) and quantified the joint effect of both loci on the disease heritability by sibling relative risk (SRR) (see supplementary Appendix 4). To ensure comparability across values of γ, we fixed the marginal relative risk at the locus of interest, and adjusted the marginal effect at the "remaining genome" locus to maintain SRR at 2, 4, and 8.

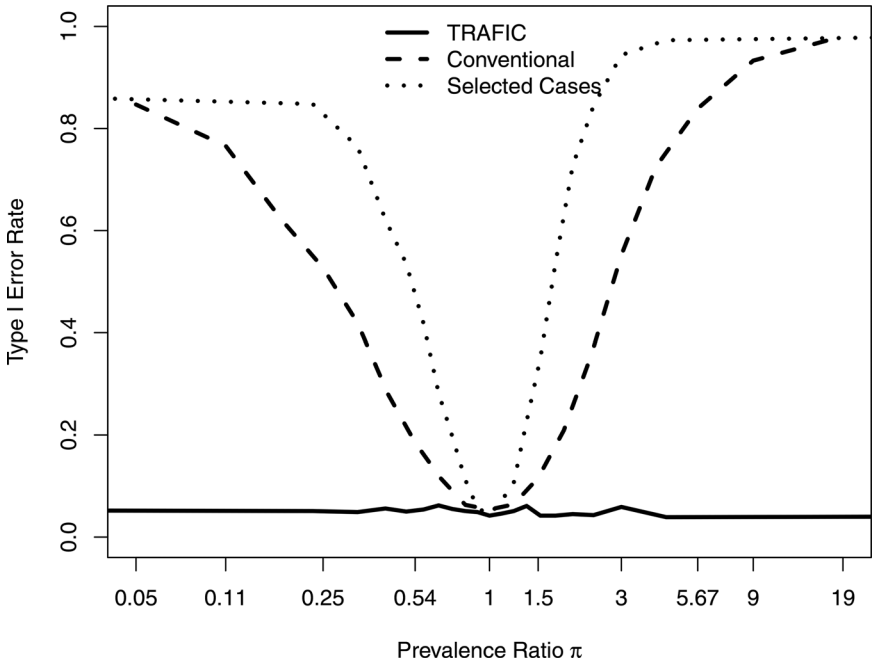

Figure 4. False-positive rate in the presence for population stratification for TRAFIC, selected cases, and the conventional case-control design. The false-positive rate is shown as a function of the prevalence ratio π between two sampled populations. Calculations are based on a summed allele frequency of 0.01 in population 1 and 0.05 in population 2, and a sample size of 2,000 individuals (1,000 sibpairs or 1,000 cases and 1,000 controls) at a significance level 0.05.

We considered a locus of interest with $f = 0.01$ and set the marginal relative risk to 2.2 for models with no interaction ($\gamma = 1$) or synergistic interaction ($\gamma > 1$), and to 2.8 for models with antagonistic interaction ($\gamma < 1$) to illustrate the effect of antagonistic interaction with reasonable power. The qualitative impact of interaction on power was independent of these specific parameter choices (results not shown).

Because the marginal effect at the locus of interest was constant, the power of the conventional case-control study was not affected by the considered interaction or by SRR. The power of TRAFIC increased with γ regardless of SRR across most interaction parameters considered (Fig. 5). For synergistic interaction, the power rose quickly with γ; the exact trajectory depended on SRR of the model. The power for models with a higher SRR increased faster for a lower γ, but the rate of increase also decreased faster for a higher SRR. Hence, models with a lower SRR reached maximal power faster. In models of antagonistic interaction ($\gamma < 1$), TRAFIC rapidly lost power with decreasing γ. This loss of power was particularly pronounced for highly heritable disease (SRR = 8). For SRR at 2, 4, and 8, TRAFIC was less powerful than the conventional design for $\gamma < 0.52$, 0.74, and 0.76, respectively (Fig. 5A). However, the power started to increase when $\gamma <$ 0.38, 0.31, and 0.26 for SRR = 2, 4, and 8, respectively. For this extreme model of antagonistic interaction, a variant that was causal in a population sample had a protective effect in a family sample. Hence, the MAF on shared chromosome regions became lower than the MAF on nonshared chromosome regions, generating power in a test for association.

Discussion

We introduce a new framework for gene-based association tests of rare variants leveraging affected sibpairs (TRAFIC). We compare the number of risk alleles located on chromosome regions shared IBD in an affected sibpair to the number of risk alleles located on chromosome regions that are not shared IBD. TRAFIC compares "cases" and "controls" within a sibpair as a matched design and is thus generally robust to population stratification. The test evaluates the null hypothesis of no association and can therefore generate a signal only in the presence of association and is powerful in the absence of linkage.

The proposed design of taking shared chromosome regions as new "cases" and nonshared chromosome regions as new "controls" can be applied to any published gene-based test. In this study, we evaluated the design for a collapsing gene-based test as the power of this test can be calculated without specifying MAF or effect size distribution of each risk variant, and it is therefore easier to obtain general conclusions. However, TRAFIC can also be applied to dispersion tests such as SKAT [Wu et al., 2011].

We calculate the power of this new method using a general model for risk variants, which is specified by the summed allele frequency of risk variants, and mean and variance of relative risk for risk variants. We compared three study designs: (1) TRAFIC, (2) the conventional design of cases and controls, and (3) a design where cases are enriched for rare variants by selecting case individuals with affected relatives

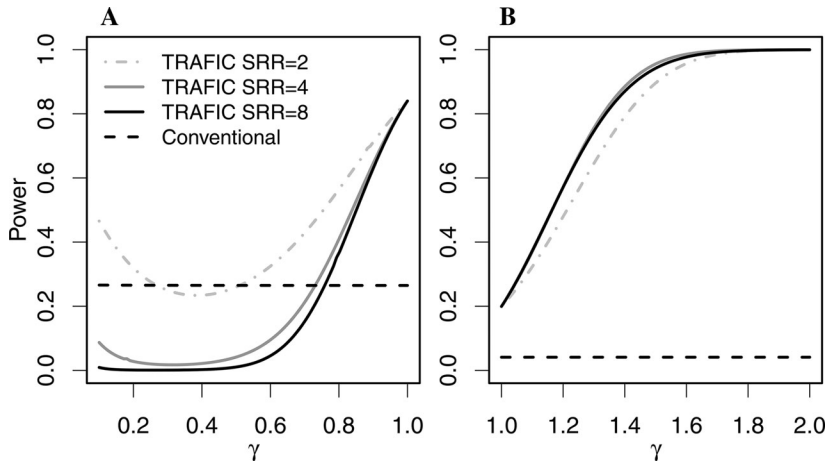

Figure 5. Analytical power of TRAFIC and the conventional case-control design under different models of gene-gene interaction. The horizontal axis displays the interaction parameter γ; gray and black lines represent different overall heritability parameterized as sibling relative risk (SRR). Panel (A) represents the result for antagonistic interaction ($\gamma < 1$); panel (B) represents the result of synergistic interaction ($\gamma > 1$). Results are shown for 2,000 individuals (1,000 sibpairs or 1,000 cases and 1,000 controls) at a significance level 2.5×10^{-6}.

assuming the same number of sequenced/genotyped samples. For diseases with prevalence ~1% and in the absence of gene-gene interaction, TRAFIC was more powerful than the conventional case-control design for variants with summed risk allele frequency less than 0.05, even though the conventional case-control design contained more independent observations. This power gain has two drivers. First, families ascertained to carry multiple affected individuals are more likely to segregate risk variants than random cases [Fingerlin et al., 2004; Peng et al., 2010; Zöllner, 2012]. Second, if such risk variants are rare, the founders of the pedigree are likely to only carry one copy. As the probability of carrying the risk variant is increased for each affected family member, this variant is more likely to be located on a shared chromosome. With increasing allelic heterogeneity, the probability for both affected siblings sharing an allele with a large effect size also rises, increasing the number of risk alleles located on shared IBD chromosome regions. Hence in the presence of allelic heterogeneity, the power of TRAFIC increased, while the power of the conventional case-control design was unchanged.

The power of a family-based design also depends on the interaction between variants at the locus of interest and the remaining genome. Sampling from families with multiple affected individuals increases the overall genetic load for all cases. Hence, if the genetic effect at the locus of interest increases with overall genetic load, the power advantage of family-based designs over population-based designs is larger than under a model of no interaction. On the other hand, if the genetic effect of risk variants at the locus of interest decreases with overall genetic load, the power in family-based designs is smaller than the power under a model of no interaction and population-based designs can be more powerful. This effect has been described before for additive gene-gene interaction, which is a special case of genetic effect at the locus of interest decreasing with overall genetic load [Helbig et al., 2013; Ionita-Laza and Ottman, 2011; Risch, 2001; Zöllner, 2012].

Moreover, TRAFFIC is generally robust to population stratification, as it compares IBD chromosome regions to nonshared chromosome regions in every sibpair thus naturally matching the genetic background of samples. This robustness can be violated in regions where one of the populations has a strong linkage signal while the other population has no evidence for linkage. However, this unlikely scenario only results in minor increase of the false-positive rate and has thus little impact on the utility of our method. As the efficacy of current methods to control for population stratification in population-based designs for rare variant tests is not clear [Mathieson and McVean, 2012, Liu, Nicolae and Chen, 2013], family-based designs may be necessary to avoid spurious association. TRAFIC achieves this robustness to stratification by using nonshared chromosome regions as controls at the cost of some reduction in power. As nonshared chromosome regions have a higher risk allele frequency than chromosome regions in population controls, a test comparing shared chromosome regions against chromosome regions from unaffected controls may be more powerful than TRAFIC. However, such a design would be very susceptible to population stratification, even more than the selected cases design shown in Figure 4.

In conclusion, we have proposed TRAFIC using affected sibpairs for testing the association between a set of rare variants and the disease phenotype. TRAFIC is more powerful than the conventional case-control design under a wide range of models while being generally robust to population stratification.

Web Resources

The R code and manual for TRAFIC can be downloaded from http://www-personal.umich.edu/~khlin/.

Acknowledgements

The authors thank Goncalo Abecasis, Michael Boehnke, and Trivellore Raghunathan for helpful discussions. This work was supported by National Institutes of Health grant HG005855.

References

Chen H, Meigs JB, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 37:196–204.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 12:628–640.

De G, Yip W, Ionita-Laza I, Laird N. 2013. Rare variant analysis for family-based design. *PloS One* 8:e48495.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004.

Fang S, Sha Q, Zhang S. 2012. Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet Epidemiol* 36:499–507.

Fingerlin TE, Boehnke M, Abecasis GR. 2004. Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. *Am J Hum Genet* 74:432–443.

Guan W, Boehnke M, Pluzhnikov A, Cox NJ, Scott LJ. 2012. Identifying plausible genetic models based on association and linkage results: application to type 2 diabetes. *Genet Epidemiol* 36:820–828.

Guo W, Shugart YY. 2012. Detecting rare variants for quantitative traits using nuclear families. *Hum Hered* 73:148–158.

Helbig I, Hodge SE, Ottman R. 2013. Familial cosegregation of rare genetic variants with disease in complex disorders. *Eur J Hum Genet* 21:444–450.

Howson JMM, Walker NM, Clayton D, Todd JA. 2009. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes Obes Metab* 11(Suppl 1):31–45.

Ionita-Laza I, Ottman R. 2011. Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 189:1061–1068.

Keith JM, McRae A, Duffy D, Mengersen K, Visscher PM. 2008. Calculation of IBD probabilities with dense SNP or sequence data. *Genet Epidemiol* 32:513–519.

Laird NM, Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7:385–394.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321.

Liu Q, Nicolae DL, Chen LS. 2013. Marbled inflation from population structure in gene-based association studies with rare variants. *Genet Epidemiol* 37:286–292.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–753.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44:243–246.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14002 people. *Science* 337:100–104.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42:30–35.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276.

Peng B, Li B, Han Y, Amos CI. 2010. Power analysis for case-control association studies of samples with known family histories. *Hum Genet* 127:699–704.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86:832–838.

Rao DC, Province MA, Leppert MF, Oberman Al, Heiss G, Ellison RC, Arnett DK, Eckfeldt JH, Schwander K, Mockrin SC, Hunt SC. 2003. A genome-wide affected sibpair linkage analysis of hypertension: the HyperGEN network. *Am J Hypertens* 16:148–150.

Risch N. 1990. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 46:222–228.

Risch N. 2001. Implications of multilocus inheritance for gene-disease association studies. *Theor Popul Biol* 60:215–220.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15:1576–1583.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, Lin X. 2012. SNP set association analysis for familial data. *Genet Epidemiol* 36:797–810.

Shugart YY, Zhu Y, Guo W, Xiong M. 2012. Weighted pedigree-based statistics for testing the association of rare variants. *BMC Genomics* 13:667.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S. 2014. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 87:604–617.

Zawistowski M, Reppell M, Wegmann M, St. Jean PL, Ehm MG, Nelson MR, Novembre J, Zöllner S. 2014. Sources of population stratification in gene-based rare variant tests identified by the joint site frequency spectrum. *Eur J Hum Genet* 22:1137–1144.

Zöllner S. 2012. Sampling strategies for rare variant tests in case-control studies. *Eur J Hum Genet* 20:1085–1091.