Analysis of case-control association studies with known risk variants

ZAITLEN, Noah, et al.

Abstract

The question of how to best use information from known associated variants when conducting disease association studies has yet to be answered. Some studies compute a marginal P-value for each Several Nucleotide Polymorphisms independently, ignoring previously discovered variants. Other studies include known variants as covariates in logistic regression, but a weakness of this standard conditioning strategy is that it does not account for disease prevalence and non-random ascertainment, which can induce a correlation structure between candidate variants and known associated variants even if the variants lie on different chromosomes. Here, we propose a new conditioning approach, which is based in part on the classical technique of liability threshold modeling. Roughly, this method estimates model parameters for each known variant while accounting for the published disease prevalence from the epidemiological literature.

Reference


DOI : 10.1093/bioinformatics/bts259
PMID : 22556366
Analysis of case–control association studies with known risk variants

Noah Zaitlen1,2,3,4,∗, Bogdan Pasaniuc1,2,3,4, Nick Patterson3, Samuela Pollack1, Benjamin Voight3,5,6, Leif Groop1, David Altshuler4,5,6, Brian E. Henderson5, Laurence N. Kolonel9, Loic Le Marchand9, Kevin Waters8, Christopher A. Haiman8, Barbara E. Stranger8,9,10, Emmanouil T. Dermitzakis11, Peter Kraft1,2,3,4 and Alkes L. Price1,2,3,4,∗

1Department of Epidemiology, 2Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, 3Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, 4Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA, 5Center for Human Genetic Research, Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, 6Departments of Genetics and of Medicine, Harvard Medical School, Boston, MA 02115, 7Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, Scania University Hospital Lund University, Malm, Sweden SE-205, 8Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA 90089, 9Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, HI 96813, 10Department of Medicine, Brigham and Women’s Hospital, Boston, MA 02115 and 11Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland CH-1211

Received on November 4, 2011; revised on March 29, 2012; accepted on April 25, 2012

1 INTRODUCTION

The NHGRI catalog of Published Genome Wide Association Studies (GWAS) (Lindert et al. 2008) lists thousands of single nucleotide polymorphisms (SNPs) associated with several hundred complex phenotypes. However, it is currently unknown how to optimally use these discovered SNPs when conducting additional GWAS. Typically, known variants are ignored and SNPs are tested independently for association via logistic regression for case–control phenotypes and linear regression for quantitative phenotypes (McCarty et al. 2008). Occasionally, known variants are used as covariates in the regression models to determine additional signals exist in the data beyond those already discovered, as in recent studies of Type 2 diabetes (Ruggieri et al. 2010). We show that for standard case–control studies neither one of these strategies, testing SNPs marginally or standard conditioning on associated variants, is optimally powered to discover new loci. Surprisingly, standard conditioning will often dramatically decrease power (Kuo and Feingold, 2011). For example, in the Welcome Trust Case Control Consortium (WTCCC), Type 1 diabetes (T1D) dataset WTCCC (2007), conditioning on a known variant on Chromosome 6 decreases the one degree of freedom (df) χ2 statistic from a logistic regression likelihood ratio test by an average of 27% at independent known associated variants on entirely different chromosomes relative to the same test without conditioning on the.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
Chromosome 6 variant. However, we find that if used properly, known variants can substantially improve study power and therefore represent an important resource in conducting future GWAS.

In this work, we thoroughly examine the use of known associated variants in the analysis of GWAS and their effects on SNPs that are completely unlinked to the known variants (i.e. on different chromosomes or distant loci). Previously, Neubauer (2009) showed that in the case of logistic regression, including a covariate increased power when it was uncorrelated to the random variable being tested, but decreased power when it was correlated. Our extensive simulations and analysis of real gene expression and case-control data indicate that for randomly ascertained individuals such as those in a cross-sectional study, the practice of standard conditioning on known variants does indeed improve the power to discover new variants (Ma et al., 2010; Neuhaus, 1998), with larger gains in power as the fraction of variance explained by the conditioned SNPs increases. However, in a balanced case–control study in which an equal number of cases and controls are ascertained based on disease status, standard conditioning on known variants significantly decreases power when the disease prevalence is low. This power loss is due to an induced non-independence between associated variants in case–control datasets. That is, SNPs that were completely uncorrelated in the population become correlated when individuals are collapsed into a case–control study design, and as predicted by Neubauer (2009), there is a corresponding loss in power due to this correlation. This is true regardless of whether the data are generated under a liability threshold model of disease or the logit model of disease assumed by logistic regression. We show that the effect of standard conditioning on known variants is a function of prevalence, sample ascertainment, and the total phenotypic variance explained by the known variants. We give full analytic derivations of the non-centrality parameter of the conditioned and unconditioned tests detailing the scenarios when each improves or diminishes power.

To address this power loss in the case–control setting, we develop a new statistic, called LTSCORE, based on the liability threshold model (Dempster and Lerner, 1955; Falconer, 1960). LTSCORE properly accounts for study design and disease prevalence while still leveraging the known associated SNPs. The basis for the improvement of our statistic is the incorporation of external prevalence information, which is readily available. The liability threshold model models individuals as having an unobserved continuous phenotype called the liability (Dempster and Lerner, 1955; Falconer, 1960). Cases are individuals whose liability exceeds some threshold while all other individuals are controls. We compute the posterior mean of the residual of the liability threshold model of disease (Dempster and Lerner, 1955; Falconer, 1960), with larger gains in power as the fraction of variance explained by the conditioned SNPs increases. However, in a balanced case–control study in which an equal number of cases and controls are ascertained based on disease status, standard conditioning on known variants significantly decreases power when the disease prevalence is low. This power loss is due to an induced non-independence between associated variants in case–control datasets. That is, SNPs that were completely uncorrelated in the population become correlated when individuals are collapsed into a case–control study design, and as predicted by Neubauer (2009), there is a corresponding loss in power due to this correlation. This is true regardless of whether the data are generated under a liability threshold model of disease or the logit model of disease assumed by logistic regression. We show that the effect of standard conditioning on known variants is a function of prevalence, sample ascertainment, and the total phenotypic variance explained by the known variants. We give full analytic derivations of the non-centrality parameter of the conditioned and unconditioned tests detailing the scenarios when each improves or diminishes power.

In practice, our disease model changes dichotomous phenotypes to continuous ones. Cases are assigned positive-valued phenotypes and controls negative-valued phenotypes. Individuals carrying a smaller number of risk alleles are given a larger phenotype. The size of these shifts are a function of SNP effect size and disease prevalence, which is not accounted for in standard logistic regression. Our approach, unlike standard logistic regression, does not suffer any loss of power when the disease prevalence is low. This is not an issue with the logit model, which may also be adapted to account for ascertainment, but with the commonly used approach of adding genetic covariates to standard logistic regression in ascertained data, without accounting for disease prevalence (see Section 4).

Results on empirical data, including a large Type 2 diabetes (T2D) case-control study and the WTCCC (2007), TID, Rhematoid Arthritis (RA), and T2D GWAS, demonstrate the pitfalls of using logistic regression with covariates as well as the power gains of LTSCORE when compared with both logistic regression with and without covariates. The gain in power is a function of prevalence and total variance explained by the known SNPs. Our method matches or outperforms conditioned or unconditioned linear or logistic regression for nearly all values of prevalence or ascertainment examined. Its performance relative to these methods will continue to increase as more variance in disease risk is explained by risk variants that are identified. We release a software package implementing LTSCORE for use in future association studies.

2 METHODS

Given a normally distributed continuous phenotype Y or a case-control phenotype Z, we wish to test candidate SNP s0 for association with the phenotype. There are K independent SNPs s1,...,sK with genotypes g1,...,gK and minor allele frequencies p1,...,pK known to be associated with the phenotype and in complete linkage equilibrium (i.e. on different chromosomes) with SNP s0. SNP s0 has genotypes g0 and minor allele frequency q0. In this work, we explore three classes of statistical tests of association: NOCOND, STDCOND and LTSCORE: NOCOND-log is logistic regression of the genotypes g0 against the phenotypes without conditioning on any known genetic covariates. STDCOND-log is logistic regression where the genotypes g1,...,gK are included as covariates. LTSCORE is linear regression applied to the posterior mean of the residual of the liability threshold model described below. NOCOND-lin and STDCOND-lin refer to linear instead of logistic regression. Each test generates a \( x^2 \) one of test statistic by performing a likelihood ratio test. Under the alternate hypothesis the effect size of \( q_0 \) is a free parameter and under the null hypothesis the effect size of \( q_0 \) is fixed at 0. The details of logistic and linear regression models are described in Supplementary Material. For reasons of simplicity, the derivations below all use linear instead of logistic regression. Linear regression is commonly used in place of logistic regression in association studies (Amstardak et al., 2008; Price et al., 2008). Furthermore, we perform simulations and experiments under both linear and logistic regression frameworks to demonstrate that the theory described below holds under both models in practice (see Section 3). The extension of these tests to recessive and dominant models is straightforward. LTSCORE is publicly available in the LITSOFT software package.

2.1 Randomly ascertained case–control phenotypes

We begin with the case of cross-sectional dichotomous phenotypes (see Supplementary Material for Continuous phenotypes). We create a dichotomous phenotype Z under a liability threshold model (Dempster and Lerner, 1955) by labeling all N individual cases when \( Y \geq \tau \) for a threshold \( \tau \), and controls
When we test the same and so the covariance will be 0. However, in a disease of low-prevalence this assumption no longer holds. That is, as the known variants explain more of the phenotype, the greater our power to discover new variants by conditioning on randomly ascertained study designs (Robinson and Jewell [109]). The non-centrality parameter for STDCOND-lin is also larger than that of NOCOND-lin in the case of randomly ascertained continuous phenotypes (see Supplementary Material).

2.2 Non-randomly ascertained case–control phenotypes

A key assumption in the derivations above is that candidates SNP $s_0$ and SNP $s_1$ are independent. However, in an ascertained case–control study, especially one for a disease of low-prevalence this assumption no longer holds. That is, candidate, SNP $s_0$ and SNP $s_1$ that are independent in the population will be correlated in the study cohort. Consider the extreme example of drawing cases from one tail of $Y$ and controls from the other. Under both a logit and liability threshold model of disease, controls will have relatively fewer copies of $s_0$ and $s_1$, while the cases will have relatively more, and so in the study, $s_0$ and $s_1$ will be correlated. This exaggerated example provides the intuition for why we see correlation in the ascertained study as we are drawing all of our cases from an extreme tail of an underlying distribution. As shown in Section 3 below, this correlation and the corresponding effects of conditioning exist under both the liability threshold model of disease and the logit model assumed by logistic regression.

The covariance between $g_0, g_1$ in the case of haploid individuals (this is easily extended to the diploid case) is

$$\text{cov}(g_0, g_1) = E[\varphi_{g_0} \varphi_{g_1}] - P_{g_0} P_{g_1}. \quad (3)$$

The expectation of the product of $g_0$ and $g_1$ is

$$\varphi_{g_0} \varphi_{g_1} = P_{g_0} = P_{g_1} = 1/2 \text{ when } Z = 0,$$

$$= P_{g_0} = P_{g_1} = 1 \text{ when } Z = 1,$$

$$= P_{g_0} = P_{g_1} = 0 \text{ when } Z = 0,$$

$$= P_{g_0} = P_{g_1} = 0 \text{ when } Z = 1,$$

$$F_0 = \text{frequency of cases in the study and } F_0 = \text{frequency of cases in the population.} \quad \text{(4)}$$

where $F_0$ is the frequency of cases in the study and $F_0$ is the frequency of cases in the population. In the case of random ascertainment, $F_0$ and $F_0$ are the same and so the covariance will be 0. However, in a disease of low prevalence, $F_0$ and $F_0$ will differ and so $s_0$ and $s_1$ will be correlated in the study due to ascertainment-induced correlation.

When we test SNP $s_0$ marginally (NOCOND-lin), the non-centrality parameter is

$$N^2 \frac{\alpha^2}{\text{var}(X)} + 2\alpha \text{var}(X) \text{cov}(g_0, g_1) \text{ when we test } s_0 \text{ conditioned on } s_1 \text{ (STDCOND-lin), the non-centrality parameter is}$$

$$N^2 \frac{\alpha^2}{\text{var}(X)} - \text{var}(X) \text{ when we test } s_0 \text{ conditioned on } s_1 \text{ (STDCOND-lin), the non-centrality parameter is}$$

$$N^2 \frac{\alpha^2}{\text{var}(X)} - \text{var}(X) \text{ where } s_0, s_1 \text{ are the expected SNP effect sizes in the study (as opposed to } s_0, s_1 \text{ the effect sizes in the population). The full details of the derivation are given in Supplementary Material (S). In the marginal case (NOCOND-lin), the shared signal of $s_0$ and $s_1$ is added to the non-centrality parameter. This implies that the power to detect $s_0$ in the marginal case is greater if there exists another SNP $s_1$ that explains a significant fraction of the variance.
the posterior mean is not normally distributed, the use of linear regression in place of logistic regression is common practice in association studies (Amundad, USP Price et al, Wallace et al). Intuitively, the above integrals have the following effect. Cases without risk alleles at other loci are assigned more extreme phenotypes than cases with risk alleles at other loci (and analogously for controls). Consider a case with no risk alleles at any of the known associated variants. To exceed the liability threshold, such an individual will require a large value $\epsilon$ relative to a case with many risk alleles at the known associated variants. Another implication of this model (as well as the relative risk model) is that the odds ratio $\beta$ will be higher when computed with cases having no known risk alleles (see note #2).

For fixed effect sizes $\beta$, as the prevalence of the disease approaches 0, the computation of $E_x(X, \beta, F, Z)$ is dominated by the threshold $m$. All of the case individuals will have approximately the same value of $E_x(X, \beta, F, Z = 1) (E_{y=1})$, and all of the controls will have approximately the same value of $E_x(X, \beta, F, Z = 0) (E_{y=0})$. Since the LTSCORE statistic is linear regression applied to $E_x(X, \beta, F, Z)$, it is equivalent to the marginal test NOCOND-log in this case of near 0 prevalence. The liability threshold model is not the only model of disease and we also derive a prevalence aware statistic from the relative risk model of disease (RCOND defined below). The RCOND model is presented in Supplementary Material S1, but we primarily focus on the LTSCORE because the relative risk model does not easily handle non-SNP covariates such as principal components.

### 2.4 Estimating $\beta$ using published prevalence

We require an estimate of the disease prevalence $P_0$ taken from the literature. In the liability threshold model, any estimates $\hat{\beta}_1$ of $\beta_1$ and $\hat{\beta}_0$ of $\beta_0$ give an expected frequency of $1/3$ in the cases and controls. Our estimate of the population minor allele frequency is

$$\hat{p}_1 = \hat{p}_0^T + \hat{p}_1 (1 - \hat{p}_0), \quad (12)$$

where $\hat{p}_0^T$ are the observed frequencies of $p_0$ in the cases and controls. Given an estimated effect size $\hat{g}_1$ of $g_1$ in the cases and controls,

$$P(Z = 1 | g_1 = 0) = 1 - (1 - \Phi(m, \hat{g}_1)(-2\hat{p}_0), \sigma_0^2)) \quad (13)$$

$$P(Z = 1 | g_1 = 1) = 1 - (1 - \Phi(m, \hat{g}_1)(-2\hat{p}_0), \sigma_1^2)) \quad (14)$$

$$P(Z = 1 | g_1 = 2) = 1 - (1 - \Phi(m, \hat{g}_1)(-2\hat{p}_0), \sigma_1^2)) \quad (15)$$

where $\Phi(x, y, z)$ is the cumulative normal distribution evaluated at $x$, with mean $y$ and variance $z$. Then

$$P(g_1 = 1 | Z = 1) = \frac{P(Z = 1 | g_1 = 1) - P(Z = 1 | g_1 = 0)}{P(Z = 1)} \quad (16)$$

and similarly for $g_1 = 1, 2$. Finally, we compute the frequency of $1/3$ in the cases given $\hat{g}_1$ and $\hat{p}_1$ as

$$\hat{p}_1^T = P(g_1 = 1 | Z = 1) + 2P(g_1 = 2) = 2P(g_1 = 1) \quad (17)$$

and similarly for controls. Using these frequencies, we can compute the squared error between the observed and expected frequencies in the cases and controls $\hat{\beta}_0 = (\hat{p}_0^T - \hat{p}_1^T)^2 + (\hat{p}_0^T - \hat{p}_1^T)^2$. We perform a binary search for 10 iterations to identify the $\hat{\beta}_0$ that minimizes $\hat{\beta}_0$. For multiple known SNPs, the $\hat{\beta}_0$ are estimated independently and combined, and only one associated SNP from any locus can be used.

### 3 RESULTS

The theory presented in Section 2 above modeled case–control phenotypes under a liability threshold model and estimated the power of linear regression with no covariates (NOCOND-lin), linear regression conditioned on known variants (STDCOND-lin) and our liability threshold model-based LTSCORE, under various ascertainment scenarios. Here, we examine the relative benefits of the three classes of statistical tests NOCOND, STDCOND and LTSCORE over simulated and real data. For NOCOND and STDCOND, we conduct most of our analyses using the logistic regression version NOCOND-log and STDCOND-log, but we have verified that NOCOND-lin and STDCOND-lin produce very similar results (see below). There are many equivalences between the logit model, the liability threshold model and the multiplicative relative risk model (see note #2). To be maximally conservative and to demonstrate that the results derived in Methods section hold for different disease models, we simulate our case–control phenotypes under a logit model. This prevents our method from having an unfair advantage due to testing the same model that generated the data. As shown below similar results were obtained when using linear instead of logistic regression and the liability threshold model instead of the logit model.

LTSCORE computes posterior mean of the residual of the liability, using liability threshold model parameters that account for disease prevalence and study design, and then uses posterior mean as input to linear regression (see Section 2). The LTSCORE parameters are estimated from published disease prevalence data. This external information, unavailable to either NOCOND-log or STDCOND-log, is the basis of the improvement of LTSCORE. We are interested in the effects of known associated SNPs on association tests for undiscovered SNPs that are in complete linkage disequilibrium (e.g. those on completely different chromosomes) with the known associated SNPs in the population. In both the simulated and real datasets below, we never condition on SNPs that are in LD with the candidate SNP. The derivations above assumed a liability threshold model of disease. However, both the STDCOND-log and NOCOND-log tests assume a logit model of disease as they are applications of logistic regression. We compare the performance of the methods by measuring the ratio of the average $x^2$ test-statistics produced by each method. This has a natural interpretation of the increase in sample size needed to obtain the equivalent power (Prichard and Pritchard, 2003). For example, if LTSCORE gives 10% increase in test-statistic over STDCOND-log, this corresponds to adding 10% more individuals to a study analyzed with STDCOND-log to achieve the power of the original study analyzed by LTSCORE.

#### 3.1 Simulated datasets

##### 3.1.1 Randomly ascertained case–control phenotypes

To examine the effect of conditioning in randomly ascertained (cross-sectional) case–control phenotypes, we generated case–control data from a logit model $P(Disease) = e^{\alpha + \beta x}$.

The affine term $\beta$ determines the prevalence $F$ of the disease in the population. To test the effects of conditioning we tested candidate SNP $x_0$ under NOCOND-log, STDCOND-log and LTSCORE. We ran 5000 simulations of 1000 cases and 1000 controls. In each simulation, there was one candidate SNP with effect size $\alpha$ and one known variant with effect size $\beta$. The fraction of variance explained with $K$ SNPs of effect size $\beta_1$ is the same as the fraction of variance explained by one SNP with effect size $\beta$. LTSCORE with $K$ SNPs of effect size $\beta_1$ produced equivalent results to using LTSCORE with one SNP of effect size $\beta_1$. (see Supplementary Material S1) and so we chose to use one SNP for simplicity. The genotypes were generated as random draws from
a binomial distribution for each simulation. We examined a range of known variant effect sizes $\beta$, a fixed candidate SNP effect size $a = 0.35$, minor allele frequencies $p_0 = p_1 = 0.2$ and $z$ (the affine term in the logit model) corresponding to a prevalence of $F = 50\%$. The results are presented in Figure 4(a). STDCOND-log always improves on NOCOND-log and the improvement is a function of the total variance explained by the known variants. LTSCORE assumes the data were generated with a liability threshold model. Despite generating data under a logit model, LTSCORE and STDCOND-log perform similarly. Reducing the prevalence $F$ (the fraction of case individuals in the population) decreases the number of cases and increases the number of controls, but both LTSCORE and STDCOND-log still outperform NOCOND-log. Results for the liability threshold-based simulation are presented in Supplementary Figure S2a, and are similar to those presented in Figure 4(a). Standard conditioning also improves power for randomly ascertained continuous phenotypes in simulations (see Supplementary Material and Fig. S1).

3.1.2 Non-randomly ascertained case-control phenotypes

We have seen that STDCOND-log improves the power to detect new variants at independent loci relative to NOCOND-log. Surprisingly, in a balanced case–control study, this is not always the case and STDCOND-log often significantly decreases the power to detect new loci. The reason for this reduction in power is the non-random ascertainment of the samples which induces a correlation between all the causal variants. The strength of the correlation between associated variants is a function of disease prevalence. The STDCOND-log test on any set of associated variants will not only remove their signal but also some of the signal from the SNP being tested. We simulated a low-prevalence case–control phenotype under a logit model as in the randomly ascertained experiments described above with $F = 0.1\%$, $a = 2.0$ and $\beta = 2.0$. We then sampled 1000 cases and 1000 controls and measured the correlation between candidate SNP $s_0$ and SNP $s_1$. The average correlation in 5000 simulations was $r^2 = 0.11 \times (\chi^2 = 220$ via Armitage trend test). We used an extreme $\beta$ to demonstrate the effect with a small number of simulations.

To examine the relative behaviors of the three classes of tests in case-control data, we simulated a case–control phenotype under a logit model as in the randomly ascertained experiments described above with $F = 0.4\%$, $a = 0.35$ and minor allele frequency $\text{MAF} = 0.20$ for both SNPs. We then sampled 1000 cases and 1000 controls. The results are presented in Figure 4(b). When $\beta$ is small, LTSCORE is nearly identical to NOCOND-log losing 0.5% in the worst case. The improvement of LTSCORE relative to NOCOND-log increases as the known variant $\beta$ explains more of the population phenotypic variance. STDCOND-log decreases in performance relative to NOCOND-log until the known variant explains at least 35% of the population phenotypic variance, at which point STDCOND-log starts to improve. However, even after the known variant explains 50% of the in study phenotypic variation, STDCOND-log achieves only 96.8% of the NOCOND-log statistic. Note that the results presented in Figure 4(b) refer to the fraction of study variance not population variance explained. Because of the ascertainment strategy, there is a significant difference between the effect sizes of the SNPs in the study and their effect size in the population. In the population, only 4.0% of individuals are cases, while in the study, 50% of individuals are cases. This skew causes the variance explained by a SNP in the population to be much smaller than the variance explained in the study (Figure 4(a)). Results for the liability threshold based simulation are similar and presented in Supplementary Figure S2a.

3.2 Non-randomly ascertained case-control data

To examine the effects of prevalence on the three tests, we fixed $\beta = 1.5$, $a = 0.35 \text{MAF} = 0.2$ and varied the disease prevalence $F$ under the same model as above. The results presented in Figure 4(a) show that the LTSCORE always outperforms STDCOND-log. STDCOND-log reduces power compared with the NOCOND-log test when the prevalence is low. However, as the prevalence increases, the study becomes more like a randomly ascertained study and the STDCOND-log test performance increases above the NOCOND-log test. LTSCORE is slightly (~2%) worse than NOCOND-log for very low prevalence (0.1%) disease and improves as the prevalence increases. This modest loss in power is removed when the data are generated under a liability threshold model (see Supplementary Fig. S3). In this case LTSCORE always outperforms NOCOND-log and STDCOND-log. It is unknown which model better represents the truth about disease.

We tested the sensitivity of our model to the misspecification of the prevalence $F$ by generating data under the same model as above for a disease with true prevalence of 3%. We tested under our LTSCORE model for a range of "estimated" prevalences. We repeated the simulation 5000 times, with 1000 cases and 1000 controls. The results are presented in Supplementary Figure S3b. Changing the estimated prevalence between 1% and 5% had a minimal effect and the performance in this case was greater than either the NOCOND-log or STDCOND-log tests. The power was greater than NOCOND-log until the specified prevalence was greater than twice the true prevalence. The maximum power was not attained at the true prevalence and we believe this is because the disease model tested (liability threshold) is different than the disease model
WTCCC T1D and RA datasets (WTCCC, 2007b). There were 1924 case–control datasets, we include the average $\chi^2$ rate of 0.050 as desired. We begin with an analysis of low-prevalence disease (T2D). We generated the data under a liability threshold model, as a function of disease prevalence. Under low-disease prevalence, there is an induced correlation between associated variants causing a sever loss of power for the STDCOND-log test relative to the NOCOND-log test. As the prevalence increases, the design is more like a random ascertained study and the STDCOND-log test outperforms the NOCOND-log test. The LTSCORE always outperforms STDCOND-log. For low-prevalence disease, LTSCORE is slightly worse than NOCOND-log and improves as the prevalence increases.

3.2 Real data sets

3.2.1 Non-randomly ascertained datasets for low-prevalence disease (T1D, RA) We begin with an analysis of low-prevalence case–control phenotypes (see Supplementary Material for real continuous phenotypes). We examined the performance of the NOCOND-log, STDCOND-log and LTSCORE statistics on the WTCCC T1D and RA datasets. There were 1924 and 1860 cases for RA and T1D respectively, and the same set of 2938 controls for the two datasets. For T1D, we used a prevalence of 0.125% (Cooper and Stroehla, 2003) and HLA SNP rs6457620 from Chromosome 6 as the known variant which explained 12.4% phenotypic variation in the study. For RA, we used a prevalence of 1% (Cooper and Stroehla, 2003) and HLA SNP rs6457620 from Chromosome 6 as the known variants which explained 7.1% phenotypic variation in the study. We filtered out all SNPs with MAF <5% and applied the NOCOND-log, STDCOND-log and LTSCORE, tests to all SNPs not found on Chromosome 6.

Although the WTCCC studies identified a relatively small number of risk loci due to limited sample size, for T1D and RA this includes HLA, a locus of large effect. The prevalences of T1D and RA are low so the expected improvement of LTSCORE relative to STDCOND-log is not expected to be large (see Section 3.1). However, these datasets demonstrate the potential for a severe loss in power of using STDCOND-log and that LTSCORE is well behaved for low-prevalence diseases. Indeed, for T1D, there was a greater than 27% drop in test statistic using STDCOND-log relative to NOCOND-log and a 4% increase using LTSCORE relative to NOCOND-log as measured by the average change in test statistic at all published GWAS variants according to the NHGRI catalog (Hindorff et al., 2009) (see Supplementary Table S1–S8).

The Q–Q plots of NOCOND-log, STDCOND-log and LTSCORE are shown in Figure 2 and serve as one means of assessing the relative performance of the methods. The significant SNPs lie at the tail of the distribution and methods with larger values at the tail are better powered. All of the test statistics had a similar $\lambda_{GC}$ and were genomic control (GC) corrected before analysis (Devlin and Roeder, 1999). On the RA dataset for example the $\lambda_{GC}$ values were 1.046, 1.047 and 1.041, for the NOCOND-log, STDCOND-log and LTSCORE tests, respectively. It is clear that STDCOND-log reduces the $\chi^2$ test statistic relative to NOCOND-log and LTSCORE in T1D (Fig. 2a) and LTSCORE (Fig. 2b). The reduction in T1D is the most dramatic because it has a very low-prevalence and the SNPs explain a larger fraction of the variance.

As another means of assessing the relative performance of the methods, we look at the test statistics of known associated variants published in the NHGRI catalog (Hindorff et al., 2009). When the known associated variant was missing from the dataset, we used the best tag as measured by $r^2$, removing any SNP where the best tag had $r^2 < 0.5$. The results are presented in Supplementary Tables S1–S2 and are analogous to the Q–Q plot results. STDCOND-log performs poorly for T1D and RA with a reduction in the sum of test statistics of roughly 27% in T1D equivalent to removing 27% of the individuals from the study (Pritchard and Preworsk, 2001). On the other hand, LTSCORE has slightly larger sum $\chi^2$ test statistics relative to NOCOND-log. We simulated 1000 case–control studies with effect sizes, prevalences and sample sizes matching the WTCCC’ studies. We generated the data under a liability threshold model and found expected gains for both studies close to 2% relative to NOCOND-log.

3.2.2 Non-randomly ascertained datasets for high-prevalence disease (T2D) We examined the performance of the NOCOND-log, STDCOND-log and LTSCORE statistics over of 6142 cases and 7403 controls genotyped at 19 known associated SNPs from the Multicohort Cohort (MEC) (African Americans, Latinos, Japanese Americans, Native Hawaiians, and European Americans) (Waters et al., 2010) and used a prevalence of 9% (Scott et al., 2009). Unfortunately, the known associated variants together explain only...
The known variants explained 1.42% and 0.64% in the original Table S3 with LTSCORE slightly outperforming NOCOND-log, \( \Delta T2D+CONTROLS \) (\( \Delta T2D+CONTROLS \)). The tail of the plots serves as an empirical measure of improvement. The ratio of cases to controls is closer to a randomly ascertained study and the STDCOND-log test suffers significant power loss. In T2D (\( \Delta T2D+CONTROLS \)) the LTSCORE outperforms the NOCOND-log test and again the STDCOND-log test suffers significant power loss. In T2D+CONTROLS (\( \Delta T2D+CONTROLS \)) LTSCORE and NOCOND-log perform similarly. LTSCORE tests applied to the WTCCC T1D, RA, T2D, and T2D+ datasets. STDCOND-log improves over NOCOND-log in this case as shown in Figure 3c. The LTSCORE method performs well in all instances, matching or outperforming each of the other tests.

4 DISCUSSION

We have shown that the practice of standard conditioning on known associated variants does not account for study design and disease prevalence potentially leading to significant power loss. This power loss is due to the induced correlation between associated variants in case-control studies. The phenomenon of higher odds ratios in cases with fewer risk alleles at other loci than in cases with more risk alleles at other loci can be viewed as a gene–gene interaction (Cordell, 2009). This is a statistical, rather than biological, interaction. By properly modeling the ascertainment and prevalence while still leveraging known associated variants, our LTSCORE statistic relative to NOCOND-log and LTSCORE in T2D 3(c). In the case of T2D+CONTROLS, the large number of controls create a study that is more similar to random ascertainment. As expected, LTSCORE improves over NOCOND-log in this case as shown in Figure 3c. The LTSCORE method performs well in all instances, matching or outperforming each of the other tests.

A recent T2D meta-analysis (Voight et al., 2010) uses the standard conditioning statistic and shows a significant gain in power. Their ratio of cases to controls is closer to a randomly ascertained study and in this case we expect STDCOND-log to outperform NOCOND-log and increase power. In addition to their beneficial study design, some of the conditioned variants are proximal to the new discoveries. Both of the elements serve to improve the power of standard conditioning. In this case we expect STDCOND-log to outperform NOCOND-log and increase power. Their ratio of cases to controls is closer to a randomly ascertained study and in this case we expect STDCOND-log to outperform NOCOND-log and increase power.
They suggest that this is due to a departure from a multiplicative
of known variants increased.

same locus. In this case, the purpose is to perform fine-mapping
variants, conditioning is also a widely used tool for SNPs in the
rest of the genome. They suggest that this is due to a departure from a
multiplicative model of disease. However, this effect may also be explained
from the non-independence of the genotypes that we described in case-
control studies. That is, some or all of the effect that was described
(log with an offset, which will perform similarly to STDCOND-log
improvement of LTSCORE is the published prevalence data and not
opposed to effect size estimation. In the case of IPW, unbiased
small and the emphasis of the community is on discovery as
Laan (2008) also offer an efficient estimator for case–control studies
(i.e. not just due to ascertainment). Although we focus on adapting
1985; Nam, 1992; Neuhaus, 1998). We derive this power loss in
considerable attention in the epidemiological literature. It is well
induced correlation. We caution that in tests for epistatic interaction
Non-HGRI [R01 HG006399 and N.Z., B.P., N.P., A.L.P.]; US National Institutes of
was supported by NIH CA63464, CA54281, U01HG004802 and
Multietnic Cohort (MEC) [R37-CA45281 to Laurence N. Kolonel].
Conflict of Interest: none declared.

REFERENCES
Baeumle,J.C. et al. (2009) Genome-wide association study and meta-analysis find that
Chang,B.L. et al. (2009) Fine mapping association study and functional analysis implicate a SNP in month at 10q11 as a causal variant for prostate cancer risk.
Duggirala,R. et al. (1997) A variance component approach to dichotomous trait linkage
Hindorff,L. et al. (2009) Potential etiologic and functional implications of genome-
Lango Allen,H. et al. (2010) Hundreds of variants clustered in genomic loci and
Monnens,G.M. et al. (2009) Genome-wide association scans for secondary traits using
Moolgavkar,S. et al. (1985) Assessing the adequacy of the logistic regression model
Case-control with known risk variants


WTCCC (2007b) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, 661–678.

