Analysis of case–control association studies with known risk variants

Noah Zaitlen1,2,3,4,∗, Bogdan Pașanuic1,2,3,4, Nick Patterson3, Samuela Pollack1, Benjamin Voight3,5,6, Leif Groop7, David Altshuler3,5,6, Brian E. Henderson8, Laurence N. Kolonel9, Loic Le Marchand9, Kevin Waters9, Christopher A. Haiman8, Barbara E. Stranger3,6,10, Emmanouil T. Dermitzakis11, Peter Kraft1,2,3,4, and Alkes L. Price1,2,3,4,∗

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

ABSTRACT

Motivation: 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 may 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.

Results: We show via simulation and application to empirical datasets that our approach outperforms both the no conditioning strategy and the standard conditioning strategy, with a properly controlled false-positive rate. Furthermore, in multiple data sets involving diseases of low prevalence, standard conditioning will often dramatically decrease power (Kuo and Feingold, 2010). For example, in the Wellcome 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

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) (Hindorff et al. 2009) 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. Occasionally, known variants are used as covariates in the regression models to determine additional signals in the data beyond those already discovered, as in recent studies of Type 2 diabetes (Hatted et al. 2011). 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, 2012). For example, in the Wellcome 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

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
the liability threshold approach is more versatile. We show that accounting of liability threshold modeling (Duggirala) is crucial for our approach and previous applications as a continuous phenotype and tested for association via linear regression. Given an individual's disease status, the disease prevalence and the known associated variants. This posterior mean is then treated as a continuous phenotype and tested for association via linear regression. Linear regression is commonly used in place of logistic regression. 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, 1958; Falconer, 1967). 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 liability, which is a free parameter and is fixed at 0. The details of logistic and linear regression models are described in Supplemental 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 (Amundsen et al., 1995; Price et al., 2006). 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 LTSOFT software package.

2 METHODS

Given a normally distributed continuous phenotype \( Y \) or a case–control phenotype \( Z \), we want to test candidate SNP \( s_0 \) for association with the phenotype. There are \( K \) independent SNPs \( s_1, \ldots, s_K \) with genotypes \( g_1, \ldots, g_K \) and minor allele frequencies \( p_1, \ldots, p_K \) known to be associated with the phenotype and in complete linkage equilibrium (e.g. on different chromosomes) with SNP \( s_0 \). SNP \( s_0 \) has genotypes \( g_0 \) and minor allele frequency \( p_0 \). In this work, we explore three classes of statistical tests of association: NOCOND, STDCOND and LTSCORE. NOCOND-log is logistic regression of the genotypes \( g_0 \) against the phenotypes without conditioning on any known genetic covariates. STDCOND-log is logistic regression where the genotypes \( g_1, \ldots, g_K \) are included as covariates. LTSCORE is linear regression applied to the posterior mean of the liability threshold model described below. NOCOND-lin and STDCOND-lin refer to linear instead of logistic regression. Each test generates a \( Z^2 \) one of test statistic by performing a likelihood ratio test. Under the alternate hypothesis the effect size of \( s_0 \) is a free parameter and under the null hypothesis the effect size of \( s_0 \) is fixed at 0. The details of logistic and linear regression models are described in Supplemental 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 (Amundsen et al., 1995; Price et al., 2006; Wallace et al., 2009). 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 LTSOFT software package.

2.1 Randomly ascertained case–control phenotypes

We begin with the case of cross-sectional dichotomous phenotypes (see Supplemental Material for Continuous phenotypes). We output a dichotomous phenotype \( Z \) under a liability threshold model (Falconer, 1967) by labeling all \( N \) individual cases when \( Y \geq t \), for a threshold \( t \), and controls
A key assumption in the derivations above is that candidates SNP $s_0$ and $s_1$ for 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 in randomly ascertained study designs (Robinson and Jewell [2000]). The non-centrality parameter for conditional case–control phenotypes based on the liability threshold model. In the case of non-randomly ascertained quantitative phenotypes, two associated variants $s_0$ and $s_1$ that are independent in the population will be correlated in the study cohort. Consider the extreme example where $FS > 2$ an SNP $s_0$ associated with the disease and having mean adjusted genotypes $g_1, g_2 < 0$. The prevalence represents a source of external data not available to the study due to ascertainment-induced correlation.

In the conditioned case (STDCOND-lin), the numerator is decreased because the shared signal of $s_0$ and $s_1$ is conditioned out. However, the denominator is also smaller since the variance of $Z$ conditioned on $s_1$ is smaller than the unconditioned variance of $Z$. The power of STDCOND-lin relative to NOCOND-lin is therefore a function of effect size, prevalence and ascertainment. Yang and colleagues [Yang et al. 2011] provide alternative derivations of the non-centrality parameter in the unconditioned case for both quantitative and case-control phenotypes based on the liability threshold model. In the case of non-randomly ascertained quantitative phenotypes, two associated variants $s_0$ and $s_1$ that are independent in the population will be correlated in the study for the reasons given above. We do not consider this case in detail in this work but note that in many cases, STDCOND-lin will reduce power significantly and we therefore caution against this statistic for non-randomly ascertained quantitative phenotypes.

### 2.3 LTSCORE statistic

We model a case-control phenotype as arising from an underlying normally distributed phenotype $\phi = \mu + \epsilon, \epsilon \sim N(0, 1)$ called the liability. Picked [1988]. Cases are those individuals with $\phi \geq 0$ and controls are those individuals with $\phi < 0$. There is a relationship between this liability scale and the relative risk model of disease described in detail previously [Eber et al. 1985] and [Hill et al. 2011]. If $F$ is the prevalence of the disease in the population then $m = \Phi^{-1}(1 - F)$, where $\Phi^{-1}(x)$ is the inverse of the cumulative normal distribution function with mean 0 and variance 1 evaluated at $x$, so that the expected proportion of individuals with $\phi \geq 0$ is $F$. A SNP $s_1$ associated with the disease and having mean adjusted genotypes $g_1, g_2 < 0$. The prevalence represents a source of external data not available to the study due to ascertainment-induced correlation.

The expectation procedure is repeated for independent known associated SNPs $s_2, \ldots, s_k$ giving a final model $\phi = \mu + \beta g_1 + \epsilon, \epsilon \sim N(0, \sqrt{1 - \var(\beta_1 g_1)})$ so that the total variance of $\phi$ is 1. Given a case-control study where SNP $s_1$ has frequency $p_1^*$, in the cases and frequency $p_1^*$ in the controls, we estimate $p_1^*$ via a method (described below) that relies on published prevalence data for the disease. This prevalence represents a source of external data not available to STDCOND-lin. The estimation procedure is repeated for independent known associated SNPs $s_2, \ldots, s_k$ giving a final model $\phi = \mu + \beta pop plus \epsilon, \epsilon \sim N(0, \sqrt{1 - \var(\beta_1 pop plus)}),$ where $X$ are the genotypes of the $K$ known SNP, and $\beta$ is a vector of the effects size $\beta_1, \ldots, \beta_k$. To use both the prevalence information and the effects of the known associated variants $s_1, \ldots, s_k$ when testing a new candidate SNP $n_0$, we compute the prior mean of the residual of the liability given the genotypes of the known variants $X,$ their effect sizes $\beta, the disease prevalence $F$ and the case-control status $Z E(\phi | X, \beta, F, Z)$:

$$E(\phi | X, \beta, F, Z = \text{Case}) = \int_{-\infty}^{\infty} x f(x | \beta, F, Z) dx,$$

$$E(\phi | X, \beta, F, Z = \text{Control}) = \int_{-\infty}^{\infty} x f(x | \beta, F, Z) dx,$$

where $\alpha^2 = 1 - \var(\beta_1 pop plus)$ the residual variance of $\phi$ after subtracting the variance from the known SNPs. The prevalence-aware liability threshold based statistic is then computed by running standard linear regression between the genotypes of the new SNP $s_0$ and the posterior mean of the residual of the liability of each individual as calculated above. Although
the posterior mean is not normally distributed, the use of linear regression in place of logistic regression is common practice in association studies.

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 \( x \) 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 at \( x \) will be higher when computed with cases having no known risk alleles. (See \textit{et al.} [2006].)

For fixed effect sizes \( \beta \), as the prevalence of the disease approaches 0, the computation of \( E[Y_i|X, \beta, F, Z] \) is dominated by the threshold \( m \). All of the case individuals will have approximately the same value of \( E[Y_i|X, \beta, F, Z=1] \) (\( E_{\text{true}} \)), and all of the controls will have approximately the same value of \( E[Y_i|X, \beta, F, Z=0] \) (\( E_{\text{controls}} \)). Since the LTSCORE statistic is linear regression applied to \( E[Y_i|X, \beta, F, Z] \), it is equivalent to the marginal test NOCOND under 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). The RCOND model is presented in the 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 \( p \) taken from the literature. In the liability threshold model, any estimates \( \hat{\beta}_1 \) of \( \beta_1 \) and \( \hat{\rho}_1 \) of \( \rho_1 \) give an expected frequency of \( s_1 \) in the cases and controls. Our estimate of the population minor allele frequency is

\[
\hat{p}_1 = \hat{p}_1^F + \hat{p}_1^C(1-F),
\]

(12)

where \( \hat{p}_1^F \) and \( \hat{p}_1^C \) are the observed frequencies of \( s_1 \) in the cases and controls.

Given an estimated effect size \( \hat{\beta}_1 \) of \( s_1 \)

\[
P(Z=1|g_1=0) = (1-\Phi(x, \hat{\beta}_1(-2\hat{\rho}_1^C), \sigma_1^2)) \tag{13}
\]

(13)

\[
P(Z=1|g_1=1) = (1-\Phi(x, \hat{\beta}_1(1-2\hat{\rho}_1^C), \sigma_1^2)) \tag{14}
\]

(14)

\[
P(Z=1|g_1=2) = (1-\Phi(x, \hat{\beta}_1(2-2\hat{\rho}_1^C), \sigma_1^2)) \tag{15}
\]

(15)

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

\[
P(g_1=0|Z=1) = P(Z=1|g_1=0)x(1-\hat{p}_1) \tag{16}
\]

and similarly for \( g_1=2 \). Finally, we compute the frequency of \( s_1 \) in the cases given \( \hat{p}_1 \) and \( \hat{\beta}_1 \) as

\[
\hat{p}_1^C = \hat{p}_1^C/(1+\hat{p}_1^C(1-F)),
\]

(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

\[
\hat{\beta}^2 = \hat{p}_1^C - \hat{p}_1^C(1-F),
\]

(17)

using 10 iterations to identify the \( \hat{\beta} \) that minimizes \( \hat{\beta}^2 \). For multiple known SNPs, the \( \hat{\beta}_i \) 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 versions 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 \textit{et al.} [2006]. 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 equilibrium (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 (\textit{e.g.} Pritchard and Przeworski, 2001). 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 \( P(\text{Disease}) = e^{\beta_1 x_1+\ldots+\beta_K x_K}/(1+e^{\beta_1 x_1+\ldots+\beta_K x_K}) \).

The affine term \( z \) 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 \( \beta \sqrt{K} \) is the same as the fraction of variance explained by one SNP with effect size \( \beta \). LTSCORE with \( K \) SNPs of effect size \( \beta \sqrt{K} \) produced equivalent results to using LTSCORE with one SNP of effect size \( \beta \) (see Supplementary Material S1) and so we chose to use one SNP for simplicity. The genotypes were generated as random draws from

\[
\text{[58-45 13/6/2012 Bioinformatics-bts258.tex]
Page: 1732 1729–1737]}

N.Zaitlen et al.
The STDCOND-log test on any set of associated variants will not improve 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 in a balanced case–control study with prevalence of 4.0% the STDCOND-log test performs similarly to NOCOND-log losing 0.5% in the worst case. The improvement of LTSCORE relative to NOCOND-log increases as the prevalence increases. This modest loss in power is removed when the data are generated under a liability threshold model (see Supplementary Material and Fig. S1). The 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 matches 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. 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 obtained at the true prevalence and we believe this is because the disease model tested (liability threshold) is different than the disease model tested with NOCOND-log.

Fig. 1. NOCOND-log, STDCOND-log, and LTSCORE simulations on case-control phenotypes. Results of NOCOND-log (logistic regression), STDCOND-log (logistic regression with covariates) and LTSCORE tests for simulated case–control datasets from a logit model. Study variance explained is the proportion of phenotypic variance explained in the study by the known association variant. For randomly ascertainment data to both the LTSCORE and STDCOND-log tests improve over the NOCOND-log tests and have similar performance. However, for non-randomly ascertainment case–control data (b) with prevalence of 4.0% the STDCOND-log test performs significantly worse than the NOCOND-log test.
There were 1924 case–control datasets, we include the average \( \chi^2 \) times. All three tests were well behaved maintaining a false positive and setting the effect size \( \alpha \) of 0.125\% (Cooper and Stroehla, 2003) and HLA SNP rs9273363 NOCOND-log, STDCOND-log and LTSCORE statistics on the continuous phenotypes). We examined the performance of the case–control phenotypes (see Supplementary Material for real

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 TID respectively, and the same set of 2938 controls for the two datasets. For TID, we used a prevalence of 0.125\% (Cooper and Stroehla, 2003) and HLA SNP rs9273363 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 TID 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 TID, 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 (Hindorff et al. 2009) (see Supplementary Tables S1–S8). The Q–Q plots of NOCOND-log, STDCOND-log and LTSCORE are shown in Figure 2a and b 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 all 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 RA (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 Table S1.3.2 and are analogous to the Q–Q plot results. STDCOND-log performs poorly for TID and RA with a reduction in the sum of test statistics of roughly 27\% in TID equivalent to removing 27\% of the individuals from the study (Pritchard and Prewer, 2003). 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 Multiethnic 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., 2001). Unfortunately, the known associated variants together explain only...
Case-control with known risk variants

Fig. 3. Q-Q plots of NOCOND-log, STDCOND-log and LTSCORE on WTCCC datasets Q-Q plots for the NOCOND-log, STDCOND-log and LTSCORE tests applied to the WTCCC T1D, RA, T2D, and T2D+ datasets. The tail of the plots serves as an empirical measure of improvement. In T1D (a), the LTSCORE outperforms the NOCOND-log test and the STDCOND-log test suffers significant power loss. In RA (b) the LTSCORE matches the performance of the NOCOND-log test and again the STDCOND-log test suffers significant power loss. In T2D (c) and T2D+CONTROLS (d) LTSCORE and NOCOND-log perform similarly. STDCOND-log improves significantly with the addition of controls, which mimics a randomly ascertainment design.

4% phenotypic variation in the study. We simulated 1000 datasets with the same sample size, a disease prevalence of 9%, and a known associated variant that accounted for 4% of the phenotypic variation. For an SNP with a minor allele frequency of 20% and an effect size on the liability scale of 0.05 (corresponding to 1.6% of the variance on the liability scale), the average improvement of LTSCORE was 3% with a standard error of 10% in the simulations. Using many SNPs of small effect produced, the same result as one SNP of large effect. The results on the MEC data are shown in Supplementary Table S4. We have shown that the practice of standard conditioning on clinical covariates is a fundamentally different problem, both because a different parameter estimation method is needed and because 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, STDCOND-log improves over NOCOND-log in this case as shown in Figure 3. 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 improves study power relative to NOCOND-log and STDCOND-log tests in case–control studies of mid-to-low prevalence diseases. This increase in power is a function of the total phenotypic variance explained by known variants and disease prevalence. The datasets examined here had either a low-prevalence or a small fraction of the variance explained and therefore we did not expect a large improvement. However, as more associated variants are discovered, the performance of LTSCORE will increase giving rise to power gains as a function of covariate effect size and disease prevalence. This approach can also be applied to clinical covariates, and in this case, an average power gain of >17% was achieved (Zaitlen et al., unpublished data). We have verified that results similar to Supplementary Table S3 are obtained when comparing genetic + clinical covariates to clinical covariates only (see Supplementary Table S4). However, conditioning on clinical covariates is a fundamentally different problem, both because a different parameter estimation method is needed and because with clinical covariates, it is often the case that samples are non-randomly ascertainment for covariate value as well as case–control status.

A recent T2D meta-analysis (Voght et al., 2012) uses the standard conditioning statistic and shows a significant gain in power. Their ratio of controls to known associated variants is closer to a randomly ascertainment 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. Yang et al. (2013) also examine the potential benefits of genome-wide conditioning in T2D. However, we believe the use of our LTSCORE statistic on these data could improve the power further.
by accounting for prevalence and ascertainment. In a recent meta-analysis of GWAS height data (Lango Allen et al. 2010), a randomly ascertainment continuous phenotype, standard conditioning revealed no new associated variants. This is due to the nature of their study design and not a contradiction of our results (see Supplementary Material S1). In their landmark TID paper, Barrett et al. 2008, find a correlation between disease risk computed from HLA SNPs and disease risk computed from 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 (correlation between MHC major histocompatibility complex risk score and non-MHC risk score) may be due to ascertainment-induced correlation. We caution that in tests for epistatic interaction (Moore and Williams 2009), this induced correlation could give rise to a spurious signal of epistatic interaction at true (marginally) associated variants.

Adjustment for informative covariates is not unique to genetics and the problem of estimation from case–control data has received considerable attention in the epidemiological literature. It is well known that regressing or stratifying on a covariate which is related to disease but not exposure of interest causes a reduction in power unless one matches on the covariate when sampling controls. Hoemmer and Lemeshow 2004, Jewell 2004, Mooijsk et al. 1985, Sund 1999, Neuhäuser 1998. We derive this power loss in terms of the liability threshold model. Neuhäuser (1998) shows the reduction in power under a logit model for any correlated covariate (i.e. not just due to ascertainment). Although we focus on adapting the liability threshold model to incorporate prevalence information, it may be possible to achieve the same result in a logistic framework. For example, if there is only one known variant, one could construct a 2 × 2 × 2 table of case–control status, candidate SNP S1 and known covariate S2. Much larger tables would be required as the number of known variants increased.

We recently proposed Monsees et al. 2009, a weighted logistic regression method (IPW) in the case of conditioning on environmental variables in case–control studies. Rose and van der Laan 2008 also offer an efficient estimator for case–control studies to account for ascertainment-induced biases. However, the focus of these works is obtaining an unbiased estimate of effect size while our concern is power (and a valid test under the null). In the case of genetic association studies, the effect sizes are generally small and the emphasis of the community is on discovery as opposed to effect size estimation. In the case of IPW, unbiased effect sizes are indeed obtained, but it under-performed relative to STDCOND-log, NOCOND-log and LTSCORE in simulations so is not considered. If the objective is to obtain unbiased effect sizes, IPW is recommended over LTSCORE. Note that the basis for the improvement of LTSCORE is the published prevalence data and not published SNP effect sizes. It is not equivalent to using STDCOND-log with an offset, which will perform similarly to STDCOND-log in the presence of ascertainment. Including an explicit interaction term in the logistic model introduces an extra df reducing the overall power.

Although this paper focuses exclusively on the use of conditioning to discover new loci that are completely unlinked to the known variants, conditioning is also a widely used tool for SNPs in the same locus. In this case, the purpose is to perform fine-mapping and better understand the genetic architecture of the known associated locus. Therefore, any drop in power due to induced correlation should not prevent researchers from using conditioning in this same-locus context. LTSCORE may improve fine-mapping efforts in some situations (see Supplementary Material). A discussion of usage and meta-analysis is given in the Supplementary Material.

ACKNOWLEDGEMENTS

The authors are grateful to D. Stram for helpful discussions.

Funding: NIH fellowship [5T32ES007142-27] N.Z.; NHGRI [R01 HG006399 and N.Z., B.P., N.P., A.L.P.]; US National Institutes of Health; National Cancer Institute [cooperative agreements U01-CA98233-07 to David J. Hunter, U01-CA98710-06 to Susan Gapstur, U01-CA98216-06 to Elio Riboli and Rudolf Kaaks, and U01-CA98758-07 to Brian E. Henderson, and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics]. The T2D study in the MEC was supported by NIH CA63464, CA54281, 1U01HG004802 and Multicentric Cohort (MEC) [R37-CA54281 to Laurence N. Kolonel].

Conflict of Interest: none declared.

REFERENCES

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.