Published in final edited form as:

Genet Epidemiol. 2014 February ; 38(2): 114–122. doi:10.1002/gepi.21785.

Leveraging Family History in Population-Based Case-Control Association Studies

Arpita Ghosh¹, Patricia Hartge², Peter Kraft³, Amit D. Joshi³, Regina G. Ziegler², Myrto Barrdahl⁴, Stephen J. Chanock², Sholom Wacholder², and Nilanjan Chatterjee^{2,*}

¹Public Health Foundation of India, New Delhi, India

²Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America

³Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America

⁴Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Abstract

Population based epidemiologic studies often gather information from study participants on disease history among their family members. Although investigators widely recognize that family history will be associated with genotypes of the participants at disease susceptibility loci, they commonly ignore such information in primary genetic association analyses. In this report, we propose a simple approach to association testing by incorporating family history information as a “phenotype.” We account for the expected attenuation in strength of association of the genotype of study participants with family history under Mendelian transmission. The proposed analysis can be performed using standard statistical software adopting either a meta- or pooled-analysis framework. Re-analysis of a total of 115 known susceptibility single-nucleotide polymorphisms, discovered through genome-wide association studies for several disease traits, indicates that incorporation of family history information can increase efficiency by as much as 40%. Efficiency gain depends on the type of design used for conducting the primary study, extent of family history, and accuracy and completeness of reporting.

Keywords

disease history in family members; first-degree relatives; meta-analysis; more powerful test for genetic association

*Correspondence to: Dr. Nilanjan Chatterjee, Chief, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive Suite 7-E146, MSC 9780, Rockville MD 20852, USA. Phone: 240-276-7319, chatternn@mail.nih.gov.

The authors have declared no conflict of interest.

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

Introduction

Most population-based epidemiologic studies exploring genetic association also have information collected at enrolment on history of disease in first-degree and sometimes even more distant relatives. Family history is expected to be related to the genotypes of participants at susceptibility loci due to Mendelian laws of transmission, but information on family history is typically omitted or not effectively utilized in standard association analyses. Recently, we and others have empirically examined such evidence of association between family history and genotypes of study participants for susceptibility loci that have already been detected through standard association analyses ignoring family history [Ghosh et al., 2012; Michailidou et al., 2013]. Standard analyses treat the disease status of study participants as the only “phenotype,” and the correlation between that phenotype and the genotype of study participants is tested using a model such as logistic regression. Sometimes investigators use family history as a covariate in a regression model, but usually avoid such analysis because conditioning on family history can unnecessarily dilute the strength of genotype-disease status associations.

Previously, the “kin-cohort” method has been proposed for using family history information for a set of genotyped individuals for estimation of absolute risk or penetrance of a disease associated with rare genetic mutation [Wacholder et al., 1998]. Efficient methods have also been proposed for genetic association analysis involving related individuals, such as those sampled in disease enriched pedigree [McPeck 2012; Stanhope and Abney 2012; Thornton and McPeck 2007]. However, such methods have been discussed primarily in the context of complex pedigrees or populations with related individuals, and application of such methods has not been evaluated for their potential to improve power of population-based case-control studies, with unrelated sampled participants, leveraging family history information.

We propose a simple approach to incorporate family history information as a “phenotype” into standard association testing. We partition the total information on association in the data into two independent test-statistics, one, the standard test-statistic based on the disease status of the participants and the other based on the disease status of the relatives, conditional on that of the participants. Under Mendelian laws of transmission, the ratio of the two underlying association parameters is expected to be approximately equal to twice the kinship coefficient. Taking advantage of that fact, we combine the two association test-statistics in an efficient manner that will serve to optimize power. The method can be implemented using a meta-analysis framework where the two test-statistics can be combined based on weights that not only depend on expected precision of the parameter estimates, but also expected strength of association relative to one another. Alternatively, the method can be implemented in a standard individual-level regression analysis by simply creating an appropriate appended dataset.

We estimate the potential gain in efficiency due to incorporation of family history into genetic association testing. By exploiting the simple form of the proposed test-statistic, we derive an analytic form for the noncentrality parameter of the proposed test-statistic and compare it to that of the standard association test-statistic. We also apply the proposed method to analyze evidence of association for a number of confirmed susceptibility single-

nucleotide polymorphisms (SNPs) for four diseases in previously conducted genome-wide association studies (GWAS), for each of which information on family history information was available. Both the theoretical and empirical comparisons indicate that direct incorporation of family history into genetic association test can substantially improve power of detection of underlying susceptibility loci without any additional cost. The proposed method, with its simple form, can also be easily adapted to perform various types of “collapsed” tests for examining evidence of association for rare variants.

Materials and Methods

Let Y^P denote the disease status (0 = unaffected, 1 = affected) and G^P the genotype at a single biallelic marker of the participant. In the following, we describe our method for incorporating first-degree family history, as information on more distant relatives is typically not available or not reliable in population-based studies. We first describe our method assuming a single first-degree relative per participant, but later comment on how disease history information on multiple relatives per participant can be handled. Let Y^R denote the disease status of the relative of the study participant. Under case-control sampling of study participants, the likelihood of observing (G^P, Y^R) conditional on Y^P , can be factorized as

$$P(G^P, Y^R | Y^P) = P(G^P | Y^P) P(Y^R | G^P, Y^P).$$

Two different tests for genetic association can be performed, one based on $P(G^P | Y^P)$, which is the standard likelihood for case-control data, and the other based on $P(Y^R | G^P, Y^P)$, which is the likelihood for observed family history given the genotype and disease status of the participant. We note that the likelihood for family-history data conditioning on participant's disease status allows any test derived from such likelihood to be independent of any associated test derived from the primary case-control likelihood.

Throughout this investigation, we assume that association of genotype with disease is modeled as a linear trend by allele count. The standard model for association for study participants can be specified as

$$\log \frac{P(Y^P = 1 | G^P)}{P(Y^P = 0 | G^P)} = \alpha^P + \beta^P G^P,$$

where β^P , the underlying log-odds-ratio, is the association parameter of interest. Prentice & Pyke (1979) showed that, when probability of disease given exposure is described by a logistic regression model, the odds ratio estimator and its asymptotic variance obtained from the retrospective likelihood are the same as those obtained from the prospective likelihood, i.e., that based on $P(Y^P | G^P)$. Hence, $\hat{\beta}^P$ and $\widehat{\text{var}}(\hat{\beta}^P)$, obtained from the prospective likelihood $P(Y^P | G^P)$, can be treated as valid estimates from the primary case-control sample.

A model for association of disease history in relative with participant's genotype given participant's disease status can be specified as

$$\log \frac{P(Y^R = 1 | G^P, Y^P)}{P(Y^R = 0 | G^P, Y^P)} = \alpha^R + \beta^R G^P + \gamma^R Y^P.$$

Thus, by fitting two separate logistic regressions, one based on the case-control status of the participant and the other based on family history, one could obtain estimates of the respective association parameters as $\hat{\beta}^P$ and $\hat{\beta}^R$. We note that as $\hat{\beta}^P$ is derived based on the primary case-control likelihood $P(G^P | Y^P)$; and $\hat{\beta}^R$ is derived based on the likelihood for observed family-history data given both the genotype and disease status of the participants: i.e., $P(Y^R | G^P, Y^P)$, the covariance of the respective likelihood score functions is zero and hence the two estimates $\hat{\beta}^P$ and $\hat{\beta}^R$ are asymptotically independent.

We now propose to combine the two estimates into a single test for association using a meta-analysis approach. We observe that for low penetrant SNPs one can approximately relate the two underlying association parameters [Thornton and McPeck 2007] as

$$\beta^R \approx 2\phi\beta^P = \frac{1}{2}\beta^P,$$

for first-degree relatives, where ϕ is the kinship coefficient between the participant and relative and is equal to $\frac{1}{4}$ for first-degree relatives (for details see Appendix A). Given two independent and unbiased estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, the best (minimum variance) linear unbiased

estimator of θ is $\frac{\frac{1}{\widehat{\text{var}}(\theta_1)}\hat{\theta}_1 + \frac{1}{\widehat{\text{var}}(\theta_2)}\hat{\theta}_2}{\frac{1}{\widehat{\text{var}}(\theta_1)} + \frac{1}{\widehat{\text{var}}(\theta_2)}}$ i.e., we combine $\hat{\theta}_1$ and $\hat{\theta}_2$ in the ratio $\frac{1}{\widehat{\text{var}}(\theta_1)} : \frac{1}{\widehat{\text{var}}(\theta_2)}$;

and the variance of the estimator is given by $\frac{1}{\frac{1}{\widehat{\text{var}}(\theta_1)} + \frac{1}{\widehat{\text{var}}(\theta_2)}}$. Using the relationship

$\beta^R \approx \frac{1}{2}\beta^P$, one can then combine the estimates $\hat{\beta}^P$ and $2\hat{\beta}^R$ in the ratio $\frac{1}{\widehat{\text{var}}(\beta^P)} : \frac{1}{4\widehat{\text{var}}(\beta^R)}$ and

thus construct an optimal meta-analysis test-statistic

$$T_{\text{meta}} = \frac{\frac{1}{\widehat{\text{var}}(\beta^P)}\hat{\beta}^P + \frac{1}{2\widehat{\text{var}}(\beta^R)}\hat{\beta}^R}{\sqrt{\frac{1}{\widehat{\text{var}}(\beta^P)} + \frac{1}{4\widehat{\text{var}}(\beta^R)}}}.$$

In standard meta-analysis, where we assume the same underlying odds ratio, the optimal weights for combining the effect size estimates are proportional to inverse variances. T_{meta} differs from the standard meta-analysis test-statistic by taking into account the expected heterogeneity in the two underlying association parameters: β^P and β^R , in constructing

optimal weights. Under the null hypothesis of no association between the marker and the trait, T_{meta} : essentially an Wald test-statistic, is asymptotically distributed as a standard normal variate.

We use asymptotic relative efficiency (ARE), defined as the limiting inverse ratio of the sample sizes needed by the two tests to attain the same power, for comparing the tests. ARE of the proposed meta-analysis test-statistic compared to the standard test-statistic based only on participants' case-control status will simply be the ratio of the respective non-centrality parameters. It can approximately be expressed as (see Appendix B)

$$\text{ARE} \approx 1 + \frac{1}{4} \frac{(1+r)}{r} (\lambda \pi^R (1 - \lambda \pi^R) + r \pi^R (1 - \pi^R)), \quad (1)$$

Where π^R denotes $P(Y^R = 1 \mid Y^P = 0)$; λ is the recurrence risk ratio, defined as

$$\lambda = \frac{P(Y^R = 1 \mid Y^P = 1)}{P(Y^R = 1 \mid Y^P = 0)};$$

and r is the expected number of controls per case under the assumed sampling design.

Although presented here for a single first-degree relative per participant, our methodology, without any further modification, can be used to incorporate information from multiple first-degree relatives based on a composite family-history variable indicating presence or absence of at least one affected first-degree relative. For relatively uncommon diseases, it is rare to have multiple first degree-relatives with the same condition in population-based studies; thus the loss of information due to such summarization of data is typically not large. However, it is also possible to generalize our method to incorporate more details on family history, such as the number of affected relatives. One could consider a multinomial logistic regression to account for the number of affected relatives. We use several approximations to derive the simplified form of the weights for combining the association signals from disease history in participants and those from disease history in relatives. It is noteworthy that it is only the power of the test, but not its validity (i.e., type I error), that depends on how accurate these approximations are, as the test-statistic is unbiased irrespective of the weights being correct or not. We simulate data under the null hypothesis of no linkage between participants' disease status and their genotype and compare the proportion of simulations in which T_{meta}^2 exceeds the 95th quantile of the χ^2 distribution to the nominal type I error level of 5%.

One always would like to report effect size estimates along with confidence intervals for GWAS findings for interpretation of results. Our method can readily provide odds-ratio estimates and confidence intervals. Our log-odds-ratio estimate would be a weighted combination of the log-odds-ratio estimate from the case-control analysis and twice the log-odds-ratio estimate from the family-history-based analysis with weights proportional to

inverse variances, i.e. $\frac{\frac{1}{\widehat{\text{var}}(\beta^P)} + \frac{1}{4\widehat{\text{var}}(\beta^R)}}{\frac{1}{\widehat{\text{var}}(\beta^P)} + \frac{1}{4\widehat{\text{var}}(\beta^R)}}$. The standard error for the estimate would be the square root of inverse of the sum of the weights, $\sqrt{\frac{1}{\frac{1}{\widehat{\text{var}}(\beta^P)} + \frac{1}{4\widehat{\text{var}}(\beta^R)}}}$, as in standard meta-analysis.

Empirical Evaluation of Power

We empirically evaluated the potential for power gain due to incorporation of family history using data on confirmed susceptibility SNPs for four common chronic diseases: breast cancer, type 2 diabetes (T2D), prostate cancer, and coronary heart disease (CHD). For each disease, a list of known common susceptibility SNPs was compiled from published reports and associations. These shortlisted SNPs were analyzed using the standard and proposed methods (see Table 2 and Appendix D). These illustrative datasets were typically smaller than those of the original GWAS that identified the susceptibility SNPs, so the majority of the SNPs did not achieve genome-wide significance in our current analyses. We, however, could utilize the confirmed SNPs identified from larger studies as “positive controls” to compare power for the two methods. For each trait, we reported the median ratio of the family-history-based and case-control log-odds-ratios across SNPs. As a measure of overall relative efficiency of the two methods, we evaluated the median ratio of the two chi-square test-statistics: proposed meta-analysis and standard case-control, over the corresponding list of susceptibility SNPs.

We restricted the analysis to a single SNP per locus when two or more SNPs within a locus were in strong linkage disequilibrium. Surrogate SNPs ($r^2 \geq 0.75$) were identified for susceptibility markers that were not typed in the study. For each SNP we fit two separate logistic regressions, one based on the case-control status of the participant and the other based on family history of the participant. For both regressions, we adjust for relevant demographic covariates; for the family-history-based logistic regression we adjust for participant’s disease status.

Results

Table 1 gives the estimated type I error at 5% level of significance for T_{meta} based on 10,000 simulations for each scenario. The empirical type I error, for each scenario, is close to the nominal level suggesting that no serious inflation occurs while assessing significance of T_{meta} .

We plot the expected ARE of the proposed method compared to standard case-control analysis, as presented in equation (1), as a function of $\pi^R = P(Y^R = 1 | Y^P = 0)$: the

prevalence of family history among controls, $\lambda = \frac{P(Y^R = 1 | Y^P = 1)}{P(Y^R = 1 | Y^P = 0)}$; which approximates the

recurrence risk ratio for rare diseases, and case-to-control ratio among participants (Fig. 1). Similar plots based on simulations (Supplementary Fig. S1) confirm our analytic results. In

general, as expected number of affected relatives of cases and controls increases, ARE of the proposed method continues to increase. However, as the proportion of cases with affected relatives, measured by $\lambda \pi^R$, increases beyond $\frac{1}{2}$, ARE drops. As the control-to-case ratio increases with everything else remaining the same, ARE of the proposed method increases. It is well known that in standard case-control analysis the marginal utility of additional controls continues to drop with increasing number of controls and becomes minimal once the control-to-case ratio reaches 4:1 [Wacholder et al., 1992]. In the proposed metaanalysis however, additional controls continue to increase the efficiency of the analysis through their contribution to family history, irrespective of the control-to-case ratio in the main study.

We note that the median ratios of the log-odds-ratios for the known SNPs associated with family history and disease status of the participants were generally close to their expected theoretical value of 0.5, except for T2D where the value was substantially attenuated towards zero (see Table 2). Analysis of known susceptibility SNPs for the four common chronic diseases using our new method (see Supplementary Table S1 and Fig. S2) confirms that incorporation of family history substantially strengthens evidence for genetic association compared to the standard case-control approach. For some individual SNPs, for example for the prostate-related SNP rs4242382, incorporation of family history lowered the magnitude of P -value by several orders. More generally, the decreases in magnitude of P -values were more modest, but were uniformly present uniformly for the majority of the SNPs across all the outcomes. Consistent with theoretical calculations, we also observed that the relative efficiency of the proposed method compared to the standard case-control method, as characterized by the median of the ratio of the corresponding chi-square association test-statistics, was largest for CHD, which was the most common condition of the four diseases we studied and had a large control-to-case ratio (see Table 2). As one would expect, the log-odds-ratio estimates from the case-control and meta-analysis were similar and the 95% confidence intervals from meta-analysis were slightly narrower than the ones from standard case-control analysis.

Discussion

Both theoretical power calculations and empirical analyses of positive-control SNPs for four common diseases indicate that incorporation of family history information directly into association testing can substantially improve power for detecting susceptibility loci in population-based case-control studies. Given that family history information often exists as part of available questionnaire data in many epidemiologic studies, the additional gain in efficiency is virtually free, apart from additional efforts needed for data harmonization. The method can be used to incorporate not only the family history information for a disease (e.g., breast cancer) from the primary study, but also from other additional studies that have been conducted to explore possibly a different outcome (e.g., prostate cancer) but might have collected family history information for the outcome of interest (e.g., breast cancer) in the primary study. Given the expected gain in efficiency especially in the large control-to-case ratio settings, such efforts may be particularly worthwhile for large consortia that may be able to share information on family history across related diseases.

We note several caveats for our method. The issue of misclassification of disease status [Ghosh et al., 2012] and misreporting of relatedness of family members are important ones. If family history data are more error prone or/and less specific compared to the primary case-control outcome under study, then there will be less gain in efficiency. In general, accuracy of reported family history information varies by characteristics of the participants, the relatives and the type of disease. Mai et al. [2011] examined the accuracy of reported family cancer history in the 2001 Connecticut Family Health Study and found that reports on breast, prostate, colorectal and lung cancer in the first- and second-degree relatives of participants had low-to-moderate sensitivity and positive predictive value, but high specificity and negative predictive value. Further, for studying association with subgroups of disease, such as a specific histologic type, or where the precise definition of case-status may be based on several relevant subphenotypes, as for a trait like CHD, disease history in relatives may be less precisely assessed than in study participants. These limitations, which are present in varying degrees in the datasets we analyzed, do not alter our conclusion that family history information adds valuable information to association analyses.

The simple framework we propose for utilizing family history information in association tests offers a flexible and versatile tool to use in a wide variety of settings. For example, a large number of methods have been proposed for testing associations with groups of rare variants—known as the burden test [Han and Pan 2010; Li and Leal 2008; Madsen and Browning 2009; Morgenthaler and Thilly 2007; Price et al., 2010], which collapse or summarize information across multiple variants, and nonburden tests that aggregate individual-variant test-statistics [Lee et al., 2012], such as, the C-alpha test [Neale et al., 2011] and the more general Sequence Kernel Association Test (SKAT) [Wu et al., 2011]. In Appendix C, we show that instead of combining the case-control and family-history-based association test-statistics via meta-analysis, one can alternatively perform a “pooled” analysis using disease status for both participants and relatives as “phenotype” in a single logistic regression. Such “pooled” analysis framework directly provides log-odds-ratio estimates along with confidence intervals and could easily be adopted for performing rare variant association tests. First, for each individual variant, a score test-statistic, incorporating both primary case-control outcome and family history information, can be obtained using the proposed “pooled” formulation. A composite group-level test-statistic can be computed by summing the per-variant score-statistics using the same types of weighting schemes as has been proposed for C-alpha and SKAT. For burden tests, we could simply replace the allele count by the burden variable in the “pooled” analysis for group-level inference.

Thornton and McPeck suggest a flexible method, M_{QLS} for utilizing family history information by optimally weighting contribution of cases and controls according to their family history [Thornton and McPeck 2007]. We show based on simulations that M_{QLS} and the proposed meta-analysis test-statistics have similar asymptotic relative efficiencies compared to the standard case-control test-statistic (Supplementary Fig. S3). The proposed method, being based on standard logistic modeling framework, have several advantages. First, the method can easily incorporate covariates, such as principal components for population stratification adjustment. Second, the method can readily provide estimates of odds ratios and confidence intervals that are typically reported from association studies for quantifying effect-sizes for detected variants. Finally, the meta-analysis framework provides

a flexible framework for pooling multiple studies some of which may be contributing association information based only on primary case-control outcome, some based only on family history of a secondary outcome and some based on both.

In conclusion, we propose a flexible approach that can be performed using standard statistical software for combining genetic association signals from the case-control status and family history information for a set of genotyped study participants. Both theoretical power calculations and re-analysis of known susceptibility SNPs for several diseases suggest that the method has the potential to accelerate discoveries in future studies and perhaps, decrease the burden of large-scale replication testing [Chanock et al., 2007].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Part of Dr. Ghosh's research and Drs. Hartge, Ziegler, Chanock, Wacholder, and Chatterjee's research was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. The breast cancer dataset was funded by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233 and U19-CA148065 to David J. Hunter, U01-CA98710 to Michael J. Thun, U01-CA98216 to Elio Riboli and Rudolf Kaaks, and U01-CA98758 to Brian E. Henderson, and Intramural Research Program of NIH/NCI, Division of Cancer Epidemiology and Genetics).

Funding support for the GWAS of Gene and Environment Initiatives in T2D was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). The human subjects participating in the GWAS derive from The Nurses' Health Study and Health Professionals' Follow-up study and these studies are supported by National Institutes of Health grants CA87969, CA55075, and DK58845. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01HG004424). The datasets used for the analyses described in this manuscript were obtained from dbGaP at [<http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap>] through dbGaP accession number [phs000091].

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022, R01HL087641, R01HL59367, and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

Appendix A: Relationship Between Underlying Association Parameters

Assuming polygenic mode of inheritance with a large number of independently segregating loci contributing in an additive fashion, we can derive the conditional distribution for the polygenic risk score of the participant given family history and participant's own disease status.

Let $Z^P = \beta^P G^P$ denote the "risk-score" associated with a participant given her/his genotype G^P . If we assume that the normal approximation $Z^P | Y^P = 0 \sim N(0, \sigma^2)$ holds, where the

centering of mean is done for convenience, then under logistic model for disease-risk,

$\frac{P(Y^P = 1 | z^P)}{P(Y^P = 0 | z^P)} \propto \exp(z^P)$, it follows that $Z^P | Y^P = 1 \sim N(\sigma^2, \sigma^2)$, as outlined below

$$\begin{aligned} p(z^P | Y^P = 1) &= \int p(z^P | Y^P = 0) \frac{p(z^P | Y^P = 1)}{p(z^P | Y^P = 0)} dz^P \\ &\propto \int p(z^P | Y^P = 0) \frac{P(Y^P = 1 | z^P)}{P(Y^P = 0 | z^P)} dz^P \\ &\propto \int p(z^P | Y^P = 0) \exp(z^P) dz^P \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(z^P - \sigma^2)^2\right\}. \end{aligned}$$

This corresponds to a disease odds ratio of $\exp(z^P)$ based on the following equation

$$\begin{aligned} OR_{Y^P}(z^P) &= \frac{P(z^P | Y^P = 1)}{P(z^P = 0 | Y^P = 1)} \bigg/ \frac{P(z^P | Y^P = 0)}{P(z^P = 0 | Y^P = 0)} \\ &= \exp(z^P). \end{aligned}$$

Moreover, it can be shown that the distribution of risk-score conditional on both the disease status of the participant and that of her/his first-degree relative can be approximated as

$$\begin{aligned} Z^P | Y^P = 0, Y^R = 0 &\sim N(0, \sigma^2) \\ Z^P | Y^P = 0, Y^R = 1 &\sim N(2\phi\sigma^2, \sigma^2) \\ Z^P | Y^P = 1, Y^R = 0 &\sim N(\sigma^2, \sigma^2) \\ Z^P | Y^P = 1, Y^R = 1 &\sim N((1 + 2\phi)\sigma^2, \sigma^2), \end{aligned}$$

where ϕ is the kinship coefficient between the participant and her/his relative (see Chatterjee et al., [2013]) This implies that

$$OR_{Y^R}(z^P | Y^P) = \frac{P(z^P = 1 | Y^P, Y^R = 1)}{P(z^P = 0 | Y^P, Y^R = 1)} \bigg/ \frac{P(z^P = 1 | Y^P, Y^R = 0)}{P(z^P = 0 | Y^P, Y^R = 0)} = \exp(2\phi z^P),$$

and therefore

$$\log OR_{Y^R}(z^P | Y^P) = 2\phi \log OR_{Y^P}(z^P).$$

If the risk-score reflected polygenic risk associated with combination of additive effects of multiple SNPs, the normal approximation would be more accurate in the above calculations. Nevertheless, the assumption provides reasonable approximation for a single common SNP as well (see Table 2).

Appendix B: Asymptotic Relative Efficiency (ARE) of Proposed Meta-analysis

The noncentrality parameter for the standard case-control association test is

$$\frac{(\beta^P)^2}{\text{var}(\beta^P)}.$$

Now

$$\text{var}_{H_0}(\widehat{\beta^P}) = \frac{1}{\frac{n_1 n_0}{(n_1 + n_0)} \text{var}(G^P)},$$

where n_1 and n_0 are numbers of cases and controls respectively in the sample leading to a noncentrality parameter of

$$\frac{n_1 n_0}{n_1 + n_0} 2(\beta^P)^2 f(1 - f),$$

where f denotes the minor allele frequency. The noncentrality parameter for T_{meta} is

$$\frac{(\beta^P)^2}{\text{var}(\widehat{\beta^P})} + \frac{(\beta^R)^2}{\text{var}(\widehat{\beta^R})}.$$

The ARE of the proposed meta-analysis test-statistic compared to the standard case-control test-statistic will simply be the ratio of the respective noncentrality parameters

$$\text{ARE} = 1 + \left(\frac{\beta^R}{\beta^P} \right)^2 \frac{\text{var}(\widehat{\beta^P})}{\text{var}(\widehat{\beta^R})}.$$

For $\text{var}_{H_0}(\widehat{\beta^R})$ we can think of $\widehat{\beta^R}$ as an inverse-variance weighted average of log-odds-ratio estimates, $\widehat{\beta_{ca}^R}$ and $\widehat{\beta_{co}^R}$, in cases and controls respectively, i.e.,

$$\widehat{\beta^R} = \frac{\frac{1}{\widehat{\text{var}}(\widehat{\beta_{ca}^R})} \widehat{\beta_{ca}^R} + \frac{1}{\widehat{\text{var}}(\widehat{\beta_{co}^R})} \widehat{\beta_{co}^R}}{\frac{1}{\widehat{\text{var}}(\widehat{\beta_{ca}^R})} + \frac{1}{\widehat{\text{var}}(\widehat{\beta_{co}^R})}}$$

Thus

$$\widehat{\text{var}}(\beta^R) = \frac{1}{\frac{1}{\widehat{\text{var}}(\beta_{ca}^R)} + \frac{1}{\widehat{\text{var}}(\beta_{co}^R)}}$$

with

$$\widehat{\text{var}}_{H_0}(\beta_{ca}^R) = \frac{1}{n_1 P(Y^R = 1 | Y^P = 1) P(Y^R = 0 | Y^P = 1) \text{var}(G^P)},$$

and

$$\widehat{\text{var}}_{H_0}(\beta_{co}^R) = \frac{1}{n_0 P(Y^R = 1 | Y^P = 0) P(Y^R = 0 | Y^P = 0) \text{var}(G^P)}.$$

$\Rightarrow$

$$\begin{aligned} \text{ARE} &= 1 + \left(\frac{\beta^R}{\beta^P} \right)^2 \frac{(n_1 + n_0)}{n_1 n_0} \\ &\quad \times (n_1 P(Y^R = 1 | Y^P = 1)(1 - P(Y^R = 1 | Y^P = 1)) \\ &\quad + n_0 P(Y^R = 1 | Y^P = 0)(1 - P(Y^R = 1 | Y^P = 0))) \\ &= 1 + \frac{1}{4} \frac{(1+r)}{r} (\lambda \pi^R (1 - \lambda \pi^R) + r \pi^R (1 - \pi^R)). \end{aligned}$$

Appendix C: “Pooled” analysis of disease status of participants and relatives

Instead of combining information across participants and relatives via meta-analysis of case-control and family-history-based test-statistics, we can do a “pooled” analysis whereby we fit a single logistic regression incorporating information on disease status for both participants and relatives and perform a single test for association. For each participant-relative pair we fit

$$\log \text{it} \begin{pmatrix} Y^P \\ Y^R \end{pmatrix} \sim \begin{bmatrix} 1 & 0 & G^P & 0 \\ 0 & 1 & \frac{1}{2} G^P & Y^P \end{bmatrix} \begin{bmatrix} \alpha^P \\ \alpha^R \\ \beta^P \\ \lambda^R \end{bmatrix},$$

and test for $H_0 : \beta^P = 0$ vs. $H_0 : \beta^P \neq 0$. The expected attenuation in strength of genetic association for relative is accounted for when we halve the contribution of participant’s genotype. Also, the conditioning on participant’s disease status allows the two entries per participant-relative pair to be independent.

Appendix D:: Details on Studies Used in Empirical Evaluation

Breast Cancer

We analyzed data on breast cancer cases and controls obtained from the National Cancer Institute's Breast and Prostate Cancer Cohort Consortium (BPC3), which brings together resources from several large prospective cohorts including women: American Cancer Society Cancer Prevention Study II [Calle et al., 2002]; European Prospective Investigation into Cancer and Nutrition [Riboli et al., 2002]; Multi Ethnic Cohort [Kolonel et al., 2004]; Nurses Health Study I and II [Colditz and Hankinson 2005]; Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial [Hayes et al., 2000]; Women's Health Study [Rexrode et al., 2000]; and Melbourne Collaborative Cohort Study [Severi et al., 2006]. BPC3 has been described in detail elsewhere [Hunter et al., 2005]. In our analysis, we used data on 16,324 white women with family history of breast cancer information to study associations with 22 previously established breast cancer susceptibility loci. We defined family history variable as positive if the participant's mother or sister was reported to have been diagnosed with breast cancer. We adjusted for age at baseline in 5-year categories, cohort and country (within EPIC) in both the case-control and family-history-based logistic regressions.

Type 2 Diabetes

We re-analyzed data on 5,739 white participants from whole-genome scans in nested case-control samples from two well-established prospective cohort studies, the Nurses' Health Study (NHS) and the Health Professional Follow-Up Study (HPFS). Details of these studies are furnished elsewhere [Colditz et al., 1997; Rimm et al., 1991]. We used 34 previously established susceptibility markers to compare performances of standard case-control analysis and meta-analysis. Family history of diabetes among first-degree-relatives was defined as at least one first degree relative (mother, father, brother, or sister) with T2D. We adjusted for gender and age in both regressions.

Prostate Cancer

We used genotype and medical history data on 6411 white participants from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) cohort included in previously conducted genome-wide association studies. Details on the genotyped study participants and design of the PLCO trial are presented elsewhere [Ghosh et al., 2012; Philip et al., 2000]. We examine associations with 30 SNPs previously established as susceptibility markers for prostate cancer. Out of 6,411 study participants, 1,151 men were genotyped as cases for a nested case-control study of prostate cancer and 738 participants reported one or more male first-degree relatives (father/brother/son) to have prostate cancer. For both the regressions we adjusted for age at enrollment, treating age as a continuous covariate, and the top five principal components to control for population stratification. For the family-history-based logistic regression we adjusted for participants' gender as well.

Coronary Heart Disease

The Atherosclerosis Risk in Communities (ARIC) Study is a prospective cohort for studying cardiovascular disease and its risk factors. The study consists of participants aged 45–64 years at baseline visit conducted between 1987 and 1989 and recruited from four U.S. communities. Details of the study have been described elsewhere [The ARIC Investigators 1989]. Using ARIC data we have investigated associations with 29 common susceptibility variants for coronary heart disease (CHD). For the case-control analysis, we define participants with a definite or probable hospitalized myocardial infarction, fatal CHD, cardiac procedure, or ECG myocardial infarction by 31 December, 2004 as cases. We restricted our analysis to 11,312 participants with information on history of CHD in parents. We adjusted for gender, race, center, and age at first visit (< 55 years) in both the regressions.

References

- Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML, Feigelson HS, Thun MJ. 2002 The American cancer society cancer prevention study II nutrition cohort. *Cancer* 94(9):2490–2501. [PubMed: 12015775]
- Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE. 2007 Replicating genotype–phenotype associations. *Nature* 447(7145):655–660. [PubMed: 17554299]
- Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J- H. 2013 Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* 45(4):400–405. [PubMed: 23455638]
- Colditz GA, Hankinson SE. 2005 The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer* 5(5):388–396. [PubMed: 15864280]
- Colditz GA, Manson JE, Hankinson SE. 1997 The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J Women Health* 6(1):49–62.
- Ghosh A, Hartge P, Purdue MP, Chanock SJ, Amundadottir L, Wang Z, Wentzensen N, Chatterjee N, Wacholder S. 2012 Assessing disease risk in genome-wide association studies using family history. *Epidemiology* 23(4):616–622. [PubMed: 22575968]
- Han F, Pan W. 2010 A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70(1):42–54. [PubMed: 20413981]
- Hayes RB, Reding D, Kopp W, Subar AF, Bhat N, Rothman N, Caporaso N, Ziegler RG, Johnson CC, Weissfeld JL. 2000 Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controll Clin Trials* 21(6):349S–355S.
- Hunter D, Riboli E, Haiman C, Albanes D, Altshuler D, Chanock S, Haynes R, Henderson B, Kaaks R, Stram D. 2005 A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat Rev Cancer* 5(12):977–985. [PubMed: 16341085]
- Kolonel LN, Altshuler D, Henderson BE. 2004 The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat Rev Cancer* 4(7):519–527. [PubMed: 15229477]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. 2012 Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91(2):224–237. [PubMed: 22863193]
- 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(3):311–321. [PubMed: 18691683]
- Madsen BE, Browning SR. 2009 A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2):e1000384. [PubMed: 19214210]
- Mai PL, Garceau AO, Graubard BI, Dunn M, McNeel TS, Gonsalves L, Gail MH, Greene MH, Willis GB, Wideroff L. 2011 Confirmation of family cancer history reported in a population-based survey. *J Natl Cancer Inst* 103(10):788–797. [PubMed: 21562245]

- McPeck MS. 2012 BLUP genotype imputation for case-control association testing with related individuals and missing data. *J Comput Biol* 19(6):756–765. [PubMed: 22697245]
- Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK. 2013 Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 45(4):353–361. [PubMed: 23535729]
- Morgenthaler S, Thilly WG. 2007 A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 615(1):28–56. [PubMed: 17101154]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011 Testing for an unusual distribution of rare variants. *PLoS Genet* 7(3):e1001322. [PubMed: 21408211]
- Philip CP, Gerald LA, Robert SB, Sandra SB, David C, Crawford ED, Ronald F, Edward PG, Fred G, Marsha AH and others. 2000 Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clin Trials* 21(6):273S–309S. [PubMed: 11189684]
- Prentice RL, Pyke R. 1979 Logistic disease incidence models and case-control studies. *Biometrika* 66(3):403–411.
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010 Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86(6):832–838. [PubMed: 20471002]
- Rexrode KM, Lee I-M, Cook NR, Hennekens CH, Buring JE. 2000 Baseline characteristics participants in the Women's Health Study. *J Womens Health Gen Based Med* 9(1):19–27. [PubMed: 10718501]
- Riboli E, Hunt K, Slimani N, Ferrari P, Norat T, Fahey M, Charrondiere U, Hemon B, Casagrande C, Vignat J. 2002 European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 5(6b):1113–1124. [PubMed: 12639222]
- Rimm EB, Giovannucci EL, Willett WC, Colditz GA, Ascherio A, Rosner B, Stampfer MJ. 1991 Prospective study of alcohol consumption and risk of coronary disease in men. *Lancet* 338(8765):464–468. [PubMed: 1678444]
- Severi G, Morris HA, MacInnis RJ, English DR, Tilley WD, Hopper JL, Boyle P, Giles GG. 2006 Circulating insulin-like growth factor-I and binding protein-3 and risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev* 15(6):1137–1141. [PubMed: 16775172]
- Stanhope SA, Abney M. 2012 GLOGS: a fast and powerful method for GWAS of binary traits with risk covariates in related populations. *Bioinformatics* 28(11):1553–1554. [PubMed: 22522135]
- The ARIC Investigators. 1989 The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Atherosclerosis* 129:687–702.
- Thornton T, McPeck MS. 2007 Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet* 81(2):321–337. [PubMed: 17668381]
- Wacholder S, Hartge P, Struwing JP, Pee D, McAdams M, Brody L, Tucker M. 1998 The kin-cohort study for estimating penetrance. *Am J Epidemiol* 148(7):623–630. [PubMed: 9778168]
- Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. 1992 Selection of controls in case-control studies: III. Design options. *Am J Epidemiol* 135(9):1042–1050. [PubMed: 1595690]
- 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(1):82–93. [PubMed: 21737059]

Figure 1.

Asymptotic Relative Efficiency (ARE) of meta-analysis incorporating family history compared to standard case-control analysis for different scenarios. ARE is plotted as a function of π^R , the proportion of controls with affected relatives, for three different values of λ , the recurrence risk ratio. The left and right panels correspond to two different values of the case-to-control ratio in the sample.

Table 1.

Simulation-based type I error estimates for T_{meta} the meta-analysis test-statistic incorporating family history, along with standard errors for different scenarios^a

	Proportion of cases with affected relatives					
	0.02	0.10	0.20	0.50	0.75	0.85
0.01	0.050 (0.0021)	0.051 (0.0021)	–	–	–	–
0.05	–	0.048 (0.0022)	0.050 (0.0022)	0.050 (0.0021)		
0.10	–	–	0.052 (0.0022)	0.049 (0.0023)	0.052 (0.0022)	0.049 (0.0021)
0.25	–	–	–	0.050 (0.0021)	0.051 (0.0021)	0.049 (0.0022)

^aFor each simulation setup, we consider 6,000 cases and 6,000 controls for a disease with population prevalence 0.3. We assume a two-allele disease model and set the minor-allele frequency at 0.25. We vary the proportions of cases and controls with affected relatives within biologically plausible ranges such that the ratio of the proportion in cases to that in controls never exceeds 10.

Table 2.

Characteristics of illustrative studies^a for four traits and corresponding efficiency gain due to incorporation of family history information in association analyses of known susceptibility SNPs

Characteristics	Breast cancer	Type 2 diabetes	Prostate cancer	Coronary heart disease
Study parameters				
Number (%) of cases	7,583 (46.5)	2,645 (46.1)	1,151 (18)	1,735 (15.3)
Number (%) of controls	8,741 (53.5)	3,094 (53.9)	5,260 (82)	9,577 (84.7)
Number (%) of cases with at least one affected relative	1,628 (21.5)	1,234 (46.6)	213 (18.5)	916 (52.8)
Number (%) of controls with at least one affected relative	1,355 (15.5)	674 (21.8)	525 (10)	4,183 (43.7)
Relatives consist of	Mother or sister	Parents or siblings	Father, brother or son	Parents
Number of known SNPs analyzed	22	34	30	29
Median ratio of family history and case-control log-odds-ratios	0.43	0.36	0.46	0.44
Overall relative efficiency ^b	1.11	1.11	1.14	1.38

^aSee Appendix D for details on studies.

^bFor each trait, overall relative efficiency is estimated as the median of the ratio of the chi-square test-statistics obtained from the meta-analysis of case-control and family-history-based association-statistics and those obtained from standard case-control analysis, over the underlying susceptibility SNPs.