Using imputed genotype data in the joint score tests for genetic association and gene-environment interactions in case-control studies

Minsun Song[1,2], William Wheeler[3], Neil E. Caporaso[2], Maria Teresa Landi[2], Nilanjan Chatterjee[2,4,5*]

[1]Department of Mathematics and Statistiscs, University of Nevada Reno, Reno, Nevada

[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland

[3]Information Management Services, Inc., Rockville, Maryland

[4]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

[5]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, Maryland

Correspondence to: Nilanjan Chatterjee (nilanjan@jhu.edu)

Abstract

Background

Genome-wide association studies (GWAS) are now routinely imputed for untyped SNPs based on various powerful statistical algorithms for imputation trained on reference datasets. The use of predicted allele count for imputed SNPs as the dosage variable is known to produce valid score test for genetic association.

Methods

In this paper, we investigate how to best handle imputed SNPs in various modern complex tests for genetic association incorporating gene-environment interactions. We focus on case-control association studies where inference in an underlying logistic regression model can be performed using alternative methods that rely on varying degree on an assumption of gene-environment independence in the underlying population. As increasingly large scale GWAS are being performed through consortia effort where it is preferable to share only summary-level information across studies, we also describe simple mechanisms for implementing score-tests based on standard meta-analysis of "one-step" maximum-likelihood estimates across studies.

Results

Applications of the methods in simulation studies and a dataset from genome-wide association study of lung cancer illustrate ability of the proposed methods to maintain type-I error rates for underlying testing procedures. For analysis of imputed SNPs, similar to typed SNPs, retrospective methods can lead to considerable efficiency gain for modeling of gene-environment interactions under the assumption of gene-environment independence.

Conclusions

Proposed methods allow valid analysis of imputed SNPs in case-control studies of gene-environment interaction using alternative strategies that had been earlier available only for genotyped SNPs.

Key words: *one-step MLE, meta-analysis, prospective likelihood, empirical-Bayes, gene-environment independence, retrospective likelihood*

Introduction

Genome-wide association studies (GWAS) are now routinely imputed for untyped SNPs with powerful imputation algorithms [Howie et al., 2009; Browning and Browning, 2009; Li et al., 2010; O'connel et al., 2016; Loh et al., 2016] up to various reference panels such as the Hapmap (The International HapMap Consortium, 2005) and the 1000 Genomes (The 1000 Genomes Project Consortium, 2010; The 1000 Genomes Project Consortium, 2012; Sudmant, P. H. et al., 2015). Standard association tests for imputed SNPs are performed using the predicted allele count as the underlying dosage variable of the association model. Many earlier fine mapping studies based on the Hapmap panel have successfully used imputation for better characterization of common susceptibility SNPs within regions initially discovered through

typed SNPs. More recently, imputation based on the 1000 Genome reference panel in existing GWAS for several traits have led to the discovery of new susceptibility loci containing uncommon or rare susceptibility variants [Guerreiro et al., 2013; Wang et al., 2014; Horikoshi et al., 2015].

The use of expected allele count for imputed SNPs as the dosage variable is known to produce valid score-test for genetic association [Marchini and Howie, 2010]. In this paper, we investigate how to best handle imputed SNPs in various modern complex tests for genetic association incorporating gene-environment interactions. In particular, we focus on case-control association studies where inference in an underlying logistic regression model can be performed using various alternative methods that rely on varying degree on an assumption of gene-environment independence in the underlying population. As increasingly large scale GWAS are being performed through consortia effort where it is preferable to share only summary-level information across studies, we also explore how these methods could be implemented in the context of meta-analysis. We study type-I error and power of alternative methods using extensive simulation studies. An application of the methods is illustrated through a re-analysis of the National Cancer Institute GWAS of lung cancer that has been imputed by the 1000 Genome reference panel.

Methods

Options for Joint-Test of Association for Genotyped SNPs

We assume the main goal of our study is to test the association of disease-status (D) with genotype status (G) of marker SNPs in the presence of a set of environmental risk factors (X)

that are known to be associated with the disease. We consider logistic regression to specify the disease-risk model in the form

$$\Pr(D=1|G,X)=\frac{\exp(\alpha+\beta_g G+\beta_x X+\beta_{gx}G*X)}{1+\exp(\alpha+\beta_g G+\beta_x X+\beta_{gx}G*X)} \qquad (1)$$

where an interaction term between G and X is incorporated to allow the effect of the genetic factor, as measured in the odds-ratio scale, to vary by the level of the environmental factors. Commonly, SNP genotypes (G) are coded as allele count assuming a linear-trend model for association with the underlying trait. More generally, genotype could be coded according to dominant, recessive or a two degree-of-freedom saturated model. A joint-test for genetic association under the above model corresponds to a global null hypothesis in the form

$$H_0:\beta_g=0 \text{ and } \beta_{gx}=0.$$

For genotyped SNPs, a multi degrees-of-freedom joint-test of association and interaction has been studied earlier [Kraft et al., 2007]. Typically, the analysis is performed based on standard prospective logistic regression analysis of case-control data.

Alternatively the analysis can be performed based on a retrospective-likelihood [Chatterjee and Carroll, 2005] that allows enhancement of power by exploitation of an assumption of gene-environment independence in the underlying population. Under gene-environment independence assumption, a case-only analysis can also be performed for inference on the logistic regression interaction parameter [Peigorsch et al., 1994], but it is not suitable for joint-testing of genetic association and interaction. The use of gene-environment independence assumption, however, can lead to serious bias in both the joint- and interaction-tests when the underlying assumption of gene-environment independence is violated [Albert et al., 2001;

Mukherjee et al., 2008; Mukherjee et al., 2012].

A third alternative for joint-testing of association and interaction is to use an empitical-Bayes type inferential procedure that allows data adaptive shrinkage between estimates obtained from the prospective and retrospective likelihoods to strike a balance between efficiency and bias incurred by gene-environment independence assumption. Extensive simulation studies have shown that methods that exploit gene-environment independence assumption, such as retrospective- or EB- method, have substantial potential to improve power for gene-environment case-control studies compared to standrd prospective logistic regression [Mukherjee et al., 2008; Mukherjee et al., 2012]. The risk of false positives due to gene-environment correlation is generally low in many realistic situations and can be further minimized using the data adaptive EB or various types of two-stage procedures [Cornelis et al. , 2012; Mukherjee et al., 2012].

Derivation of Score-Tests

A major advantage of score-test, compared to Wald- or Likelihood-ratio test (LRT), is that it only requires imputation under the null model of no association and thus can easily incorporate expected dosage returned by popular imputation algorithms. Further for the analysis of less common and rare variants, score-tests may have more robust properties than Wald test or LRT as the number of cases or/and controls can be sparse in variant genotype categories.

Suppose that data consist of (D_u, X_u, G_u), $u = 1,...,n$ where D_u, X_u, and G_u, respectively, denote the disease status, environmental exposure, and SNP-genotype status for subject u. Let $Z = (1, X)$ and $W = (G, G * X)$ denote a partitioning of the design matrix associated with the

"nuisance" parameters, $\eta = (\alpha, \beta_x)$, and the parameters of interest, $\theta = (\beta_g, \beta_{gx})$, respectively, for the underlying logistic regression model.

Prospective (PT) Method

The standard prospective likelihood of case-control data is derived as

$$L_P = \prod_{u=1}^{n_0+n_1} \Pr(D_u \mid G_u, X_u).$$

Under the prospective-likelihood, the score-function for θ is given by

$$S_\theta^P = \sum_{u=1}^{n_0+n_1} \{W_u D_u - E_\eta(W_u D_u \mid G_u, X_u)\}.$$

Under the null hypothesis of no association,

$$S_{\theta_0}^P = \sum_{u=1}^{n_0+n_1} Z_u \{G_u D_u - G_u E_\eta(D_u \mid X_u)\}. \qquad (2)$$

The maximum likelihood estimator (MLE) of the nuisance parameters η under the null model can be estimated by fitting the null model

$$\Pr(D = 1 \mid X) = \frac{\exp(\alpha + \beta_x X)}{1 + \exp(\alpha + \beta_x X)}. \qquad (3)$$

The multivariate score-test-statistic can be computed as

$$T^p = (S_{\theta_0}^P)^T (V_{S_{\theta_0}}^P)^{-1} (S_{\theta_0}^P),$$

where $V_{S_{\theta_0}}^P$ is the variance-covariance matrix for the score-vector accounting for uncertainty associated with estimation of the nuisance parameters. One can estimate $V_{S_{\theta_0}}^P$ using the efficient information matrix in a model-based fashion or using the empirical variance-

covariance matrix of the associated influence function to achieve robustness against mis-specification of the null model.

Retrospective (RT) Method

The retrospective likelihood for case-control data is given by

$$L_{\mathbb{R}} = \prod_{u=1}^{n_0+n_1} \Pr(X_u, G_u \mid D_u).$$

It has been long known that inference for the parameters of interest under underlying logistic regression model is equivalent under the retrospective and prospective likelihoods for case-control data when no assumption is made about joint distribution of the underlying risk-factors, i.e. G and X in our example [Prentice and Pyke, 1979]. However, if an assumption of gene-environment independence is invoked, then more efficient inference is possible under the retrospective likelihood. In particular, Chatterjee and Carroll [2005] have previously shown that under the assumption of gene-environment independence, but without any further restriction on the distribution of X, inference under retrospective likelihood can be made using a "profile-likelihood" of the form

$$L^*_{\mathbb{R}} = \prod_{u=1}^{n_0+n_1} \Pr^*(D_u, G_u \mid X_u, R_u = 1)$$

where the conditioning $R = 1$ is introduced to indicate the selection mechanisms of subjects into the sample under the case-control sampling scheme. Derivation of L^*_R requires specification of population genotype frequencies, either using two parameters under a general multinomial model or using a single parameter under the assumption of Hardy Weinberg Equilibrium (HWE). Thus, for the retrospective likelihood, we expand the nuisance parameter

vector as $\eta^* = (\eta, \gamma)$ so that the nuisance parameters include both parameters of the disease-risk and genotype frequency models.

The score-function for association parameters of interest for the retrospective likelihood can be derived as

$$S_\theta^R = \sum_{u=1}^{n_0+n_1} \{W_u D_u - E_{\eta^*}^*(W_u D_u \mid X_u)\},$$

where $E_{\eta^*}^*$ denotes expectation with respect to the joint probability distribution of D and G given X and R=1.

Under the null,

$$S_{\theta_0}^R = \sum_{u=1}^{n_0+n_1} Z_u \{G_u D_u - E_\gamma(G_u) E_\eta(D_u \mid X_u)\},$$

which differs from the corresponding score-vector (equation (2)) is derived under the prospective likelihood only in the way the expectation is derived in the second term. In particular, under the retrospective likelihood, the expectation term is evaluated under the assumption of gene-environment independence while the prospective likelihood does not require any such assumption. Under the null hypothesis, the parameters of the null model (3) can be estimated using standard prospective logistic analysis since the MLE under the retrospective- and prospective- likelihoods are the same as we allow the distrbution of non-genetic risk-factors to be completely unspecified. Further, under null, MLE associated with genotype frequency model γ can be obtained from the pooled sample of the cases and controls. The multivariate-score test can now be derived under the retrospective likelihood following the same-steps as that for described for the propsective likelihood (see

Supplementary Methods Section 1.2 for complete details).

EB Procedure

Implementation of the original EB procedures requires parameter estimates from the prospective- and retrospective-likelihood methods. The estimate itself cannot be directly derived from a likelihood and thus derivation of a score-test for this procedure is not straightforward. As an alternative, we propose a "score-type" test that could maintain some of the advatnages of the score-tests as described earlier and yet allow combining inference from the prospective- and retrospective-likelihoods in a data adaptive fashion to balance between bias and efficiency. We first note that any score-test can be written in the form of a Wald-like test-statistic

$$T = (S_{\theta_0})^T (V_{S_{\theta_0}})^{-1} (S_{\theta_0}) = (\hat{\theta}_0)^T V_{\hat{\theta}_0}^{-1} (\hat{\theta}_0)$$

where $\hat{\theta}_0 = V_{S_{\theta_0}}^{-1} S_{\theta_0}$ can be viewed as a one-step MLE starting from the null parameter value θ_0 =0. Thus, taking advantage of the above Wald-like representation of score-test, we propose an EB score-type test in the form

$$T^{EB} = (\hat{\theta}_0^{EB})^T V_{\hat{\theta}_0^{EB}}^{-1} (\hat{\theta}_0^{EB})$$

where the corresponding EB estimates and associated variance-covariance matrix are obtained by combining the one-step MLE estimates derived from the prospective- and retrospective-likelihood (see Supplementary Methods Section 1.3) using formulae analogous to those described for the original EB procedure [Mukherjee and Chatterjee, 2008].

Derivation of the PT, RT, and EB methods under a more general setting that allows accounting for additional covariates in the model is given in Supplementary Methods.

Handling Imputed Genotype Data

Once the forms of the score-tests are derived with observed genotyped data, handling imputed genotype data for all the procedures is relatively straightforward as it simply involves replacing G_u by $\hat{G}_u$, the expected value of genotype dosage taking into account predicted probabilities of different genotype values returned by the imputation algorithm. It is noteworthy how imputed genotype data are handled differentially in the prospective- and retrospective- score functions. Under the prospective-likelihood, the score-function for imputed genotype data takes the form

$$S_{\theta_1}^{P} = \sum_{u=1}^{n_0+n_1} Z_u\{\hat{G}_u D_u - \hat{G}_u E_\eta(D_u \mid X_u)\},$$

where the imputed genotype-dosage variable contributes to both terms of the left hand side of the equation. In contrast, under the retrospective-likelihood, the score function for imputed genotype data takes the form

$$S_{\theta_1}^{R} = \sum_{u=1}^{n_0+n_1} Z_u\{\hat{G}_u D_u - E_\gamma(G_u)E_\eta(D_u \mid X_u)\}$$

where the imputed genotype-dosage variable contributes only to the first term of the equation. The genotype frequency parameters, required in derivation of the retrospective-score function, can be estimated from imputed genotype data based on overall predicted genotype counts observed in the pooled sample of cases and controls. Derivations of efficient-information matrices and empirical variance-covariance matrices for the score-vectors follow the same steps as those for observed genotype data for each of the respective procedures. Finally, the

derivation of the one-step MLEs and score-type test using the EB procedure follows the same steps as those described for observed genotype data.

Analysis of NCI GWAS of Lung Cancer

We analyzed data from a GWAS of lung cancer generated at the National Cancer Institute. The dataset included 5713 cases and 5736 controls from four different study sites (Table 1). The samples were originally genotyped using a combination of Illumina GWAS platforms and were imputed using the 1000 Genomes Phase 2 reference panel using IMPUTE2 software [Howie et al., 2009]. The details of the studies can be found in several previous publications [Landi et al.,2009]. We evaluated the performance of the different methods in evaluating joint association of lung cancer with SNP-genotypes and genotype-by-smoking interactions. We derived score test under a logistic regression model where SNP genotypes were coded assuming additive effects. For modeling the effect of smoking status, recorded as current, former or never, we used two dummy variables. The resulting joint tests for association and interaction had three degrees of freedom. We also examined the two degree-of-freedom score-tests associated with only the interaction parameters of the model, but the underlying p-values were derived under the global null hypotheses of absence of both association and interactions. Figure 1 shows the quantile-quantile (Q-Q) plots for the interaction-only tests for the application of the PT, RT, and EB methods (left panel) and the joint- tests under the PT, RT, and EB methods together with the test for main effect of G of the model without interaction (right panel), which were restricted to the analysis of ~5.3 million SNPs such that MAF > 0.05, the imputation quality reported to have info measure $I_A \geq 0.5$, and the p-values from all the seven

tests are available. As the patterns are generally similar for the model-based and empirical variance estimators, we only show results using the former method. In general, the Q-Q plot associated with the interaction-only tests aligns close to the diagonal line indicating that all the methods are maintaining type-I error well. The Q-Q plot neither shows any strong upward curvature near low p-values that could be indicative of the presence of many strong interactions in the data.

In contrast, the Q-Q plot for the main-effect-only and joint tests of association and interaction clearly shows a strong upward curvature near the tail of the distribution. This pattern is largely driven by SNPs in the chromosome 15q25.1 region which are previously shown to be strongly associated with the risk of lung cancer (See Supplementary Figure S1 for plots after removal of this region). SNPs in this region, which contains multiple nicotine receptor genes, have been shown to be associated with both risk of lung cancer [Amos et al., 2008; Thorgeirsson et al., 2008] and smoking intensity [Thorgeirsson et al., 2008;Saccone et al.,2010]. However, no SNPs in this region has been reported to be associated with smoking status even in studies with extremely large samples size (N>100K) [The Tobacco and Genetics Consortium, 2010]. Thus it is interesting that in this region (x-axis p-value < 10^{-4}), the RT and EB method, both of which exploit an assumption of independence of genotype and smoking status, consistently produce lower p-values for the SNPs than those from the main-effect only test and the joint-test under the PT method. It appears that, in this data, although gene-environment interactions by themselves are not identifiable at a high significance level, proper accounting for these effects using efficient methods are enhancing the detection of underlying signals captured by the joint test.

We also evaluated the performance of different methods including SNPs with lower MAF (MAF=0.01-0.05). In this setting, we observe that the Q-Q plots for the methods that used sandwich variance-covariance estimators were highly inflated indicating systematic problem with type-I error rate control. The problem could be traced to small sample bias of the sandwich standard errors because of small number of non-smoking cases (N=355) who also carried variant genotype for rare SNPs in our study. When we combined non-smokers and former-smokers together to a single category, the bias went away (data not shown).

Simulation Studies

We generated data on a binary environmental exposure variable which is assumed to follow Bernoulli (0.5) and be independently distributed of G. We simulated SNP genotype (G) assuming HWE and minor allele frequency (MAF) value of 0.3 or 0.05. Given the values of G and X, we generated the binary disease outcomes for individuals from the logistic regression model (1). We chose $\beta_x = \log(1.5)$ or $\beta_x = \log(2)$ to allow the association of X with D to be either modest or strong, respectively. For evaluation of type-I error, we assumed no genetic association, i.e. both $\beta_g = 0$ and $\beta_{gx} = 0$. For evaluation of power, we set $\beta_{gx} = \log(1.2)$ and $\beta_g = \log(1)$ or log(1.05). In all simulations β_0 was set such that an overall disease rate in the underlying population is about 5%. For evaluating type-I error and power, we simulated 500,000 and 1,000 datasets, respectively, with each set consisting of 5,000 controls and 5,000 cases.

To evaluate validity and power of the methods when the SNP of interest may not be genotyped, we simulated haplotypes which consist of the SNP of interest and neighboring SNPs in linkage

disequilibrium (Table 2). In one setting (left panel), the variant of interest was common (MAF=0.3) and could be imputed with high accuracy ($R^2 = 0.8$) based on genotypes of the neighboring SNPs (See Stram [2004] for R^2). In the other setting (right panel), the variant of interest was less common (MAF= 0.05) and could be predicted with moderate accuracy ($R^2 = 0.5$) based on the genotype status of the neighboring SNPs. Assuming HWE in the general population, multi-locus genotypes of individuals were generated from simulted haplotypes. We analytically evaluated the conditional probability for genotype at the SNP of interest for each configuration of genotypes at the neighboring SNPs. For simulating "imputed dosage" for the SNP of interest, we simulated the genotype data for the neigboring SNPs first and then assigned predicted genotype probabilities for the SNP of interest using the known conditional probabilities. For analysis of each simulated data, we pretended that only the predicted probabilities, and not the actual genotypes, were available at the SNP of interest.

We implemented each of the PT, RT, and EB score tests using either a model-based or an empirical variance-covariance estimator. However, in evaluation of the EB procedure, due to the lack of a model-based formula, the covariance between prospective and retrospective estimators was always evaluated based on empirical covariance of the underlying influenced functions. For each method, we evaluated the performance of both the joint- and interaction-only tests. In general, simulation studies show that the proposed methods perform well in maintaining type-I error both at modertate (α= 0.05) and stringent (α= 0.0001) significance levels (Figure 2). In some scenarios, the RT method, when implemented with the sandwich variance estimator, showed a slight inflation over the nominal significance level. Across all the

scenarios, the EB method was conservative, a pattern that has been reported earlier for analysis of typed SNPs and has been traced to the use of a conservative variance estimator [Mukherjee et al., 2012]. Employing the PT, RT, and EB methods on typed SNPs shows consistent results (Supplementary Figure S2).

Simulation studies of power (Table 3) suggest that relative performance of three different methods was similar for untyped SNPs as has been reported for typed SNPs in earlier studies [Mukherjee et al., 2012]. In particular, the RT method had the maximum power, the PT method has the minimum power and the EB procedure performed in between. All methods lost power to a similar degree for the analysis of untyped SNPs compared to the analysis of the same SNP had it been typed. The use of model-based versus sandwich variance estimators did not have much effect in power for any of the methods (Supplementary Table S1).

Discussion

In summary, we propose various types of score tests for genetic association and gene-environment interactions for analysis of case-control studies. Similar to standard tests for genetic association, in these methods, imputed genotype data for untyped SNP could be handled by simply substituting genotype values with predicted dosage that could be available from popular imputation software.

The prospective and retrospective score-tests are derived directly from the underlying likelihoods for case-control studies. We derived the score-test for the EB procedure using

underlying one-step maximum-likelihood estimates of parameters obtained from the prospective- and retrospective-likelihoods. The one-step MLEs can also be used to perform multivariate meta-analysis of the parameters across studies and then derive various test-statistics based on meta-analyzed parameter estimates and their variance-covariance matrices. In our implementation of all the methods in the R software package CGEN (https://www.bioconductor.org/packages/release/bioc/html/CGEN.html) we allow returning of the one-step MLEs to facilitate meta-analysis.

Both analysis of simulated and real datasets suggest that the proposed methods can generally control type-I error rates, but small sample bias could arise in the presence of sparse genotype-by-exposure cells, especially if sandwich variance estimators are used in some of these methods. Simulation studies of power show that the relative performances of the PT, RT, and EB procedure are quite similar for the analysis of untyped and typed SNPs. Although not studied directly, it can be anticipated that in the presence of gene-environment correlation in the population, the relative performance of these methods for their ability to control type-I error would also be similar as has been reported in earlier studies [Mukherjee et al., 2012] for typed SNPs.

Although all the methods are valid for both continuous and categorical exposures, our numerical studies only involve categorical exposures. Future studies are needed to investigate performance of the proposed methods in the presence of continuous exposure and model mis-specification. It has been noted before that if the model for association of the disease with a

continuous exposure is mis-specified, then the test for genetic associations and interaction could be biased due to underestimation of variance of target parameters under the mis-specified model [Tchetgen and Kraft, 2011]. We focus on test for genetic association and interactions using single genetic markers. Further studies are also merited how these methods could be extended for derivation of gene-level aggregate tests of genetic associations and interactions.

Acknowledgement

Research of Minsun Song, Maria Teresa Landi, Neil Caporaso, and Nilanjan Chatterjee was supported by the intramural program of the Nationcal Cancer Institute.

Software link

The proposed method has been implemented in open source software, available at https://www.bioconductor.org/packages/release/bioc/html/CGEN.html .

References

Albert, P. S., Ratnasinghe, D., Tangrea, J. , Wacholder, S. 2001. Limitations of the case-only design for identifying gene–environment interactions. Am J Epidemiol. **154**, 687–693.

Amos, C.I., Wu, X., Broderick, P. et al. 2008. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. **40**(5), 616-622.

Browning, B. L., Browning, S.R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. **84**, 210–223.

Chatterjee, N., Carroll, R. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. **92**(2), 399-418.

Chen, Y.H., Chatterjee, N., Carroll, R.J. 2009. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. J Am Stat Assoc. **104**, 220–233.

Cornelis, M.C., Tchetgen, E.J., Liang L., Qi, L., Chatterjee, N., et al. 2012. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol. **175**,191–202.

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C. et al. 2013. TREM2 variants in Alzheimer's disease. N Engl J Med. **368**, 117–127.

Horikoshi, M., Mgi, R., van de Bunt, M. et al. 2015. Discovery and fine-mapping of glycaemic

and obesity-related trait loci using high-density imputation. PLoS Genet. **11**: e1005230

Howie, B. N., Donnelly, P., Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide assocation studies. PLoS Genetics. **5**: e1000529.

International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. Nature. **467**, 52-58.

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., Gauderman, W. J. 2007. Exploiting gene–environment interaction to detect genetic associations. Hum Hered. **63**, 111–119.

Lanndi, M . T. et. al. 2009. A genome-wide assocation study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. **85**, 679-691.

Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. **34**, 816–834 .

Lin, D.Y., Tang, Z. Z. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. **89**:354-367.

Loh, P. R., Palamara, P.F. & Price, A.L. 2016. Fast and accurate long-range phasing and imputation in a UK Biobank cohort. Nat. Genet. http://dx.doi.org/10.1038/ng.3571.

Marchini, J., Howie, B. 2010. Genotype imputation for genome-wide association studies. Nature Rev Genet. 11, 499–511.

Mukherjee, B. et al. 2008. Tests for gene–environment interaction from case–control data: a novel study of type I error, power and designs. Genet Epidemiol. **32**, 615–626

Mukherjee, B. and Chatterjee, N. 2008. Exploiting gene-environment independence in analysis of case-control studies : An empirical Bayes approach to trade-off between bias and efficiency.

Biometrics. **64**(3), 685-694.

Mukherjee, B., Ahn, J., Gruber, S.B., Chatterjee, N. 2012. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. Am J Epidemiol. **175(3)**:177–190.

O'Connell, J. et al. 2016. Haplotype estimation for biobank-scale data sets. Nat. Genet. http://dx.doi.org/10.1038/ng.3583.

Piegorsch, W. W., Weinberg, C. R. , Taylor, J. A. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. Stat Med. **13**, 153–162.

Prentice, R. L., and Pyke, R. 1979. Logistic Disease Incidence Models and Case-Control Studies. Biometrika. **66**: 403–411.

Saccone, N. L. et al. 2010. Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. PLoS Genet. **6**, e1001053 .

Servin, B., Stephens, M. 2007. Imputation-based analysis of association studies: candidate genes and quantitative traits. PLoS Genetics. **3**, e114.

Sudmant, P. H. et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature **526**, 75–81.

Stram, D.O. 2004. Tag SNP seclection for association studies. Genetic Epidemiology **27**, 365–374.

Tchetgen, E.J.T. and Kraft, P. 2011. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. Epidemiology **22**(2):257-261.

The Tobacco and Genetics Consortium. 2010. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat. Genet. **42**, 441–447.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature **467**, 1061–1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1092 human genome. Nature. **491**, 56-65.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature. **437**(7063), 1299–1320.

Thorgeirsson, T. E. et al. 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature **452**, 638–642.

Wang, Y., McKay, J.D., Rafnar, T. et al. 2014. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. **46**(7):736–741.

Tables and Figures

	Cases					Controls				
cohort	ATBC	CPSII	EAGLE	PLCO	All	ATBC	CPSII	EAGLE	PLCO	All
complete data	1732	695	1978	1814	5713	1270	674	1978	1814	5736

Table 1. Distribution of cases and controls by cohorts in NCI GWAS.

GWAS, genome-wide association studies.

Table 2. Haplotypes and their frequencies used for conducting simulation studies in scenario where underlying causal SNP is untyped and is assumed to be imputed based on neighboring genotyped SNPs . "U" and "T" indicate the untyped and typed SNP positions, respectively.

MAF*=0.3, $R^2 = 0.8$ #		MAF*=0.05, $R^2 = 0.5$ #	
UTTTT	Frequency	UTTTT	Frequency
10011	0.2530	00111	0.3800
10101	0.0128	01110	0.2350
10111	0.0342	01111	0.2900
00101	0.2374	11001	0.0456
00111	0.2233	11111	0.0044
01110	0.2393	00001	0.0450

MAF, minor allele frequency.

*Minor allele frequency of the untyped causal SNP

Obtained by fitting multivariate regression of genotype at the causal SNP on the genotypes of the neghboring SNPs

Table 3. Simulation results for power of the joint- and interaction-tests for different procedures under various scenarios. For MAF=0.3, power is shown for nominal significance levels of 0.0001 and 0.001 for the joint- and interaction- tests, respectively. For MAF as 0.05, power is shown for the nominal significance level of 0.05 for both types of tests. In all settings, power was evaluated under an interaction odds-ratio=1.2. Results are shown when causal SNP is typed (bottom panels) or imputed (top panels). Variance is estimated based on information matrix.

Untyped SNPs								
MAF	β_x	β_g	**PT-joint**	**RT-joint**	**EB-joint**	**PT-int**	**RT-int**	**EB-int**
0.3	log(1.5)	log(1)	0.419	0.713	0.518	0.243	0.574	0.406
0.05	log(1.5)	log(1)	0.342	0.498	0.404	0.231	0.389	0.309
0.3	log(1.5)	log(1.05)	0.822	0.948	0.873	0.233	0.565	0.415
0.05	log(1.5)	log(1.05)	0.558	0.692	0.603	0.220	0.411	0.302
0.3	log(2)	log(1)	0.446	0.744	0.568	0.234	0.498	0.376
0.05	log(2)	log(1)	0.366	0.517	0.430	0.223	0.356	0.307
0.3	log(2)	log(1.05)	0.818	0.952	0.871	0.212	0.483	0.367
0.05	log(2)	log(1.05)	0.599	0.733	0.643	0.230	0.392	0.312
Typed SNPs								

MAF	β_x	β_g	PT-joint	RT-joint	EB-joint	PT-int	RT-int	EB-int
0.3	log(1.5)	log(1)	0.604	0.874	0.724	0.349	0.728	0.559
0.05	log(1.5)	log(1)	0.443	0.613	0.521	0.271	0.491	0.4
0.3	log(1.5)	log(1.05)	0.943	0.987	0.954	0.335	0.724	0.551
0.05	log(1.5)	log(1.05)	0.694	0.804	0.709	0.286	0.495	0.360
0.3	log(2)	log(1)	0.647	0.887	0.731	0.321	0.645	0.499
0.05	log(2)	log(1)	0.469	0.645	0.54	0.286	0.443	0.373
0.3	log(2)	log(1.05)	0.942	0.992	0.955	0.333	0.662	0.531
0.05	log(2)	log(1.05)	0.707	0.83	0.744	0.284	0.466	0.379

PT-joint , prospective joint test ; RT – joint, retrospective joint test ; EB-joint, empirical Bayes joint test; PT-int, prospective interaction test; RT-int, retrospective interaction test; EB-int, empirical Bayes interaction test; SNP, single nucleotide polymorphism; MAF, minor allele frequency.

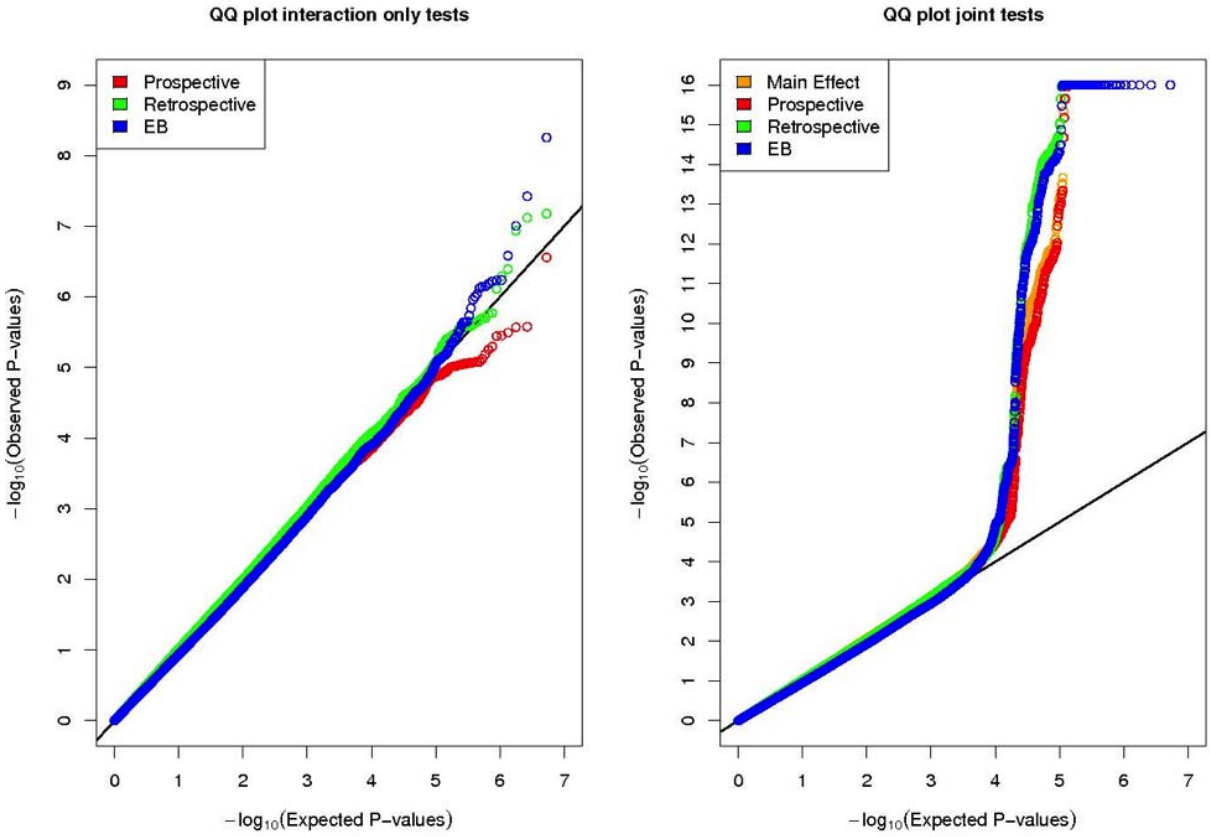
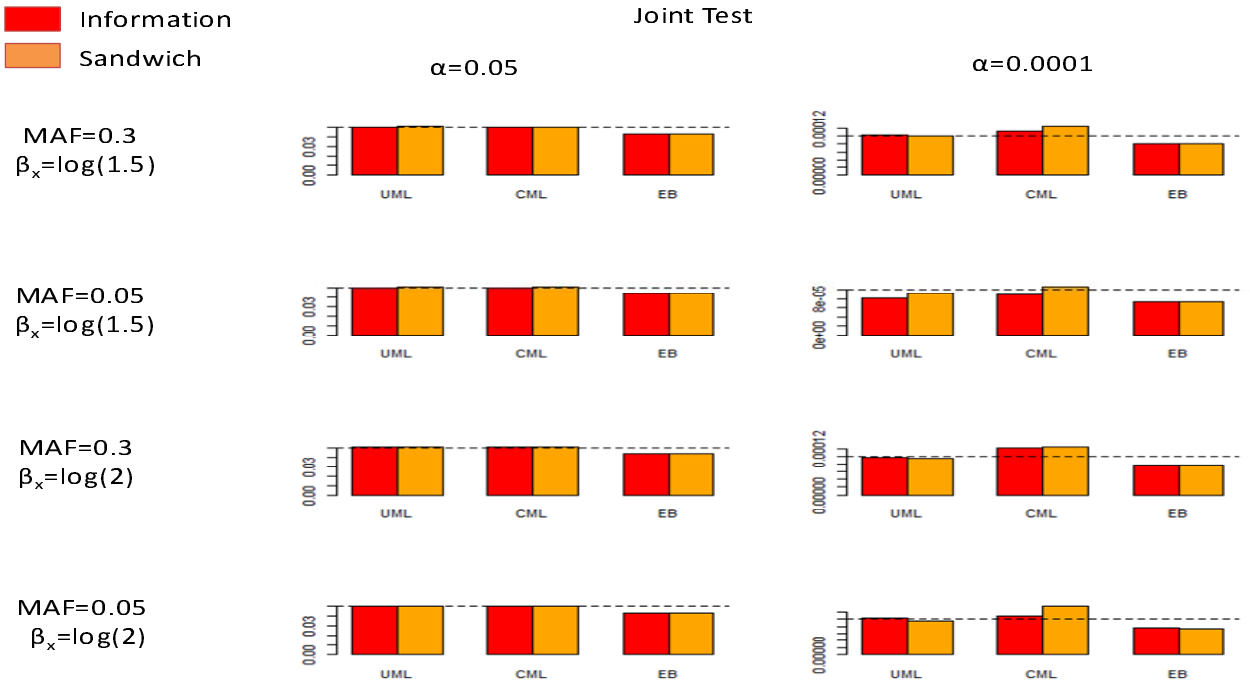


Figure 1. Quantile-quantile plots for the interaction-only, joint tests and tests for main effect of G in the analysis of National Cancer Institute Lung Cancer GWAS. Tests for associations are performed between risk of lung cancer and each of approximately 5.3 million common SNPs accounting for interactions with smoking status (never, former, and current) of the individuals. Each curve pertains to SNPs such that MAF > 0.05, the imputation quality reported to have info measure $I_A \geq 0.5$, and the p-values from all the seven tests are available. GWAS, genome-wide association studies; SNP, single nucleotide polymorphism; MAF, minor allele frequency.



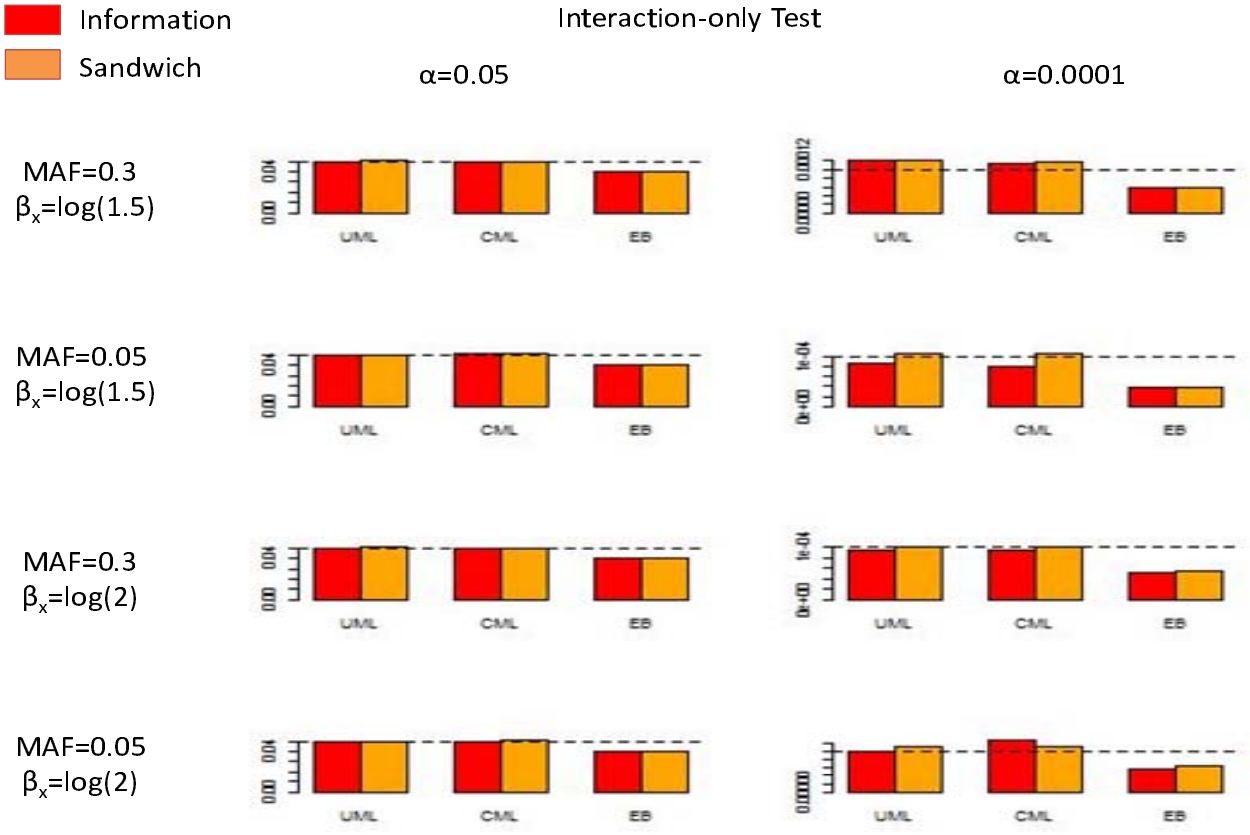


Figure 2. Simulation results for type-I error for different procedures for testing untyped SNPs. The nominal significance levels are 0.05 and 0.0001. (top panels) Joint test, (bottom panels) Interaction-only test. Red and orange pertain to information-based variance estimator and sandwich variance estimator, respectively. MAF, minor allele frequency.