Published in final edited form as:
Genet Epidemiol. 2010 December ; 34(8): 803–815. doi:10.1002/gepi.20527.

Analysis of Untyped SNPs: Maximum Likelihood and Imputation Methods

Y.J. Hu and **D.Y. Lin***
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina

Abstract

Analysis of untyped single nucleotide polymorphisms (SNPs) can facilitate the localization of disease-causing variants and permit meta-analysis of association studies with different genotyping platforms. We present two approaches for using the linkage disequilibrium structure of an external reference panel to infer the unknown value of an untyped SNP from the observed genotypes of typed SNPs. The maximum-likelihood approach integrates the prediction of untyped genotypes and estimation of association parameters into a single framework and yields consistent and efficient estimators of genetic effects and gene-environment interactions with proper variance estimators. The imputation approach is a two-stage strategy, which first imputes the untyped genotypes by either the most likely genotypes or the expected genotype counts and then uses the imputed values in a downstream association analysis. The latter approach has proper control of type I error in single-SNP tests with possible covariate adjustments even when the reference panel is misspecified; however, type I error may not be properly controlled in testing multiple-SNP effects or gene-environment interactions. In general, imputation yields biased estimators of genetic effects and gene-environment interactions, and the variances are underestimated. We conduct extensive simulation studies to compare the bias, type I error, power, and confidence interval coverage between the maximum likelihood and imputation approaches in the analysis of single-SNP effects, multiple-SNP effects, and gene-environment interactions under cross-sectional and case-control designs. In addition, we provide an illustration with genome-wide data from the Wellcome Trust Case-Control Consortium (WTCCC) [2007].

Keywords

case-control studies; cross-sectional studies; genome-wide association studies; genotype; haplotype; Hardy-Weinberg equilibrium; retrospective likelihood

INTRODUCTION

The rapid improvement of high-throughput genotyping technology and the precipitous drops of genotyping cost have led to the widespread use of genome-wide association studies (GWAS) in elucidating the genetic basis of complex human diseases. Because the current genotyping platforms assay only a small fraction of single nucleotide polymorphisms (SNPs) in the human genome, many disease-susceptibility loci will inevitably be untyped (i.e. not genotyped). Thus, it is highly desirable to conduct association analysis at untyped SNPs. Such analysis can help localize causal variants and facilitate selection of SNPs for follow-up studies. Such analysis also allows investigators to compare or combine results from studies that use different genotyping chips. As untyped SNPs are not measured on any

*Correspondence to: D.Y. Lin, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420. lin@bios.unc.edu.

study subject, the missing information cannot be recovered from the study data alone. Fortunately, the linkage disequilibrium (LD) structure observed in an external reference panel, such as the HapMap [The International HapMap Consortium, 2005], can be used to predict untyped variants from typed variants.

A variety of methods have been developed for the statistical analysis of untyped SNPs. In particular, de Bakker et al. [2005#x0005D;, Nicolae [2006], and Zaitlen et al. [2007] used the haplotype frequencies of tag SNPs to estimate the allele frequencies of the untyped SNP for cases and controls. This strategy, although simple and intuitive, is not statistically efficient and is confined to case-control comparisons without environmental factors.

In a similar spirit of tagging, Lin et al. [2008] proposed a likelihood-based method for the analysis of untyped SNPs in case-control studies with or without environmental factors. The likelihood integrates the study data and external data while reflecting the biased sampling nature of the case-control design. This method yields consistent and efficient estimators of genetic effects and gene-environment interactions, and the variance estimators fully account for the uncertainties in inferring the unknown variants.

A simpler approach is to impute the unknown values of the untyped SNPs for each subject and then use the imputed values in a downstream association analysis. Statistically speaking, this two-stage strategy, which is called single imputation in the missing data literature, is less satisfactory than the maximum likelihood because of its bias and inefficiency [Little, 1992]. However, single imputation has a practical advantage: once the missing data are imputed, the association analysis can be readily carried out (for any traits and study designs) in standard software packages.

Given the operational convenience of (single) imputation and the statistical optimality of maximum likelihood, comprehensive comparisons of these two approaches are sorely needed. This article provides such comparisons under cross-sectional and case-control designs. We expand the approach of Lin et al. [2008] to encompass both cross-sectional and case-control studies. In addition, we develop a tagging-based imputation strategy. We establish the theoretical properties of the proposed imputation method and conduct extensive simulation studies to evaluate the performance of the imputation and maximum-likelihood methods in testing/estimating genetic effects and gene-environment interactions. We apply the two methods to the GWAS data from the Wellcome Trust Case-Control Consortium (WTCCC).

METHODS

IMPUTATION

Suppose that we are interested in a particular untyped SNP, whose genotype is denoted by G_u. Let Y denote the phenotype of interest, which can be quantitative or qualitative. Also, let $\mathbf{X}$ denote a set of environmental factors. We characterize the effects of genetic and environmental factors on the phenotype through the conditional density function $P_{\alpha,\boldsymbol{\beta},\xi}(Y|G_u, \mathbf{W})$, where $\mathbf{W}$ consists of $\mathbf{X}$ and the genotypes of the typed SNPs, and α, $\boldsymbol{\beta}$ and ξ pertain to the intercept, regression parameters, and nuisance parameters (e.g. error variance), respectively. (If we are interested in the marginal effect of G_u, then $\mathbf{W}$ is an empty set.) We formulate $P_{\alpha,\boldsymbol{\beta},\xi}(Y|G_u, \mathbf{W})$ through a generalized linear regression model with linear predictor $\alpha+\boldsymbol{\beta}^T z(G_u, \mathbf{W})$, where $z(G_u, \mathbf{W})$ is a vector-function of G_u and $\mathbf{W}$ under a particular mode of inheritance. We assume the additive mode of inheritance in this article, although all the formulas can be easily modified to accommodate other modes of inheritance. For a quantitative trait, we specify the linear regression model:

$$Y=\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u,\mathbf{W})+\varepsilon,$$

where ε is zero-mean normal with variance σ^2. For a binary trait, it is natural to use the logistic regression model:

$$\Pr(Y=1|G_u,\mathbf{W})=\frac{e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u,\mathbf{W})}}{1+e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u,\mathbf{W})}}. \quad (1)$$

We use the LD information from a reference panel to select a set of (M−1) typed SNPs that provides the most accurate prediction of the untyped SNP, where M is a small number, which is set to five in this article. The accuracy of prediction is measured by R^2 of Stram [2004]. The M-locus genotype G consists of G_u and G_t, where G_t is the genotype of the (M −1) typed SNPs. Suppose that the M SNPs have a total of K haplotypes. For k = 1, …, K, let h_k denote the kth haplotype, and π_k denotes the frequency of h_k. Assume that the Hardy-Weinberg equilibrium (HWE) holds. For a reference panel of $\tilde{n}$ trios, the likelihood for $\boldsymbol{\pi}$ = $(\pi_1, \ldots, \pi_K)^{\mathrm{T}}$ is

$$L_R(\boldsymbol{\pi})=\prod_{j=1}^{\tilde{n}}\sum_{(h_k,h_l,h_{k'},h_{l'})\sim G_j}\pi_k\pi_l\pi_{k'}\pi_{l'}, \quad (2)$$

where G_j = (GF_j, GM_j, GC_j) is the genotype data for the jth trio with the M-locus genotypes GF_j, GM_j and GC_j for the father, mother and child, respectively, and (h_k, h_l, $h_{k'}$, $h_{l'}$) ~ G_j means that (h_k, h_l) is compatible with GF_j, ($h_{k'}$, $h_{l'}$) is compatible with GM_j, and (h_k, $h_{k'}$), (h_k, $h_{l'}$), (h_l, $h_{k'}$), or (h_l, $h_{l'}$) is compatible with GC_j.

By maximizing $L_R(\boldsymbol{\pi})$ given in Equation (2) via the EM algorithm, we obtain the maximum-likelihood estimator (MLE) $\tilde{\boldsymbol{\pi}}$ = $(\tilde{\pi}_1, \ldots, \tilde{\pi}_K)^{\mathrm{T}}$. Assuming that the haplotype frequencies are the same between the study population and the external panel, we can estimate the probability distribution of G_u from the observed values of G_t for each study subject according to the formula

$$\Pr(G_u=g|G_t;\tilde{\boldsymbol{\pi}})=\frac{\sum_{(h_k,h_l)\sim(G_t,G_u=g)}\tilde{\pi}_k\tilde{\pi}_l}{\sum_{g'=0,1,2}\sum_{(h_k,h_l)\sim(G_t,G_u=g')}\tilde{\pi}_k\tilde{\pi}_l},\quad g=0,1,2, \quad (3)$$

where (h_k, h_l) ~ (G_t, G_u = g) means that (h_k, h_l) is compatible with (G_t, G_u = g). We use this (estimated) probability distribution to impute the unknown value of G_u, either as the expected count (i.e. dosage) or the most likely value of G_u. We replace the unknown values of G_u by the imputed values for all study subjects to create a “complete” data set, which is then analyzed by standard regression methods.

In the Appendix A, we prove that the above imputation method yields a valid test of the null hypothesis H_0:β_G = 0 under the linear predictor $\alpha+\beta_{G_u}G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}$, where β_{G_u} and $\beta_{\mathbf{W}}$ pertain to the effects of G_u and $\mathbf{W}$, respectively, provided that G_t is independent of Y conditional on $\mathbf{W}$. This result holds for both cross-sectional and case-control studies, even when the

reference panel and the study sample are drawn from different underlying populations. However, the estimator of β_{G_u} is generally biased with underestimated variance when $\beta_{G_u} \neq 0$, and type I error may not be properly controlled for other hypotheses.

MAXIMUM LIKELIHOOD

Let H denote the diplotype associated with the M-locus genotype G. We write $H = (h_k, h_l)$ if the diplotype consists of haplotypes h_k and h_l. In the previous subsection, we formulate the effects of G and $\mathbf{X}$ through the conditional density function $P_{\alpha,\boldsymbol{\beta},\xi}(Y|G_u, \mathbf{W})$, where $\mathbf{W}$ consists of G_t and $\mathbf{X}$. In this subsection, we represent the same regression model in the form of $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathcal{G}(h_k, h_l), \mathbf{X})$, where $\mathcal{G}(h_k, h_l)$ denote the genotype G induced by the diplotype (h_k, h_l). We assume that H and $\mathbf{X}$ are independent.

Let n denote the total number of study subjects. For $i = 1, \ldots, n$, let Y_i, G_{ti}, and $\mathbf{X}_i$ denote the values of Y, G_t, and $\mathbf{X}$ on the ith subject. For a cross-sectional study, the likelihood for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\pi}$ takes the form

$$L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{(h_k,h_l)\sim G_{ti}} P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\mathcal{G}(h_k, h_l), \mathbf{X}_i)\pi_k\pi_l \tag{4}$$

where $(h_k, h_l) \sim G_{ti}$ means that the diplotype (h_k, h_l) is compatible with genotype G_{ti}.

For case-control studies, we assume the logistic regression model given in (1) with the linear predictor $\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathcal{G}(h_k, h_l), \mathbf{X})$. Because the sampling is conditional on the case-control status, the likelihood takes the retrospective form $\prod_{i=1}^{n} P(G_{ti}, \mathbf{X}_i|Y_i)$. If there are no environmental factors and the disease is rare, then this likelihood becomes

$$L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \frac{\sum_{(h_k,h_l)\sim G_{ti}} e^{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathcal{G}(h_k,h_l))}\pi_k\pi_l}{\sum_{k,l} e^{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G(h_k,h_l))}\pi_k\pi_l} \tag{5}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$. In the presence of $\mathbf{X}$, the retrospective likelihood involves the unknown distribution of $\mathbf{X}$, which is high-dimensional. We eliminate the distribution of $\mathbf{X}$ by the profile-likelihood approach [Lin and Zeng, 2006] and replace (5) with the following profile likelihood:

$$L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \frac{\sum_{(h_k,h_l)\sim G_{ti}} e^{Y_i\{\mu+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathcal{G}(h_k,h_l),\mathbf{X}_i)\}}\pi_k\pi_l}{\sum_{k,l,y} e^{y\{\mu+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathcal{G}(h_k,h_l),\mathbf{X}_i)\}}\pi_k\pi_l},$$

where $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, μ is an unknown constant, and the summation in the denominator is taken over $k, l = 1, \ldots, K$ and $y = 0,1$.

The likelihood that combines the study data and the reference panel is $L_C(\boldsymbol{\theta},\boldsymbol{\pi}) = L_S(\boldsymbol{\theta},\boldsymbol{\pi})L_R(\boldsymbol{\pi})$, where $L_R(\boldsymbol{\pi})$ is given in Equation (2). We maximize this combined likelihood via the Newton-Raphson algorithm. We set the initial value of $\boldsymbol{\pi}$ at $\tilde{\boldsymbol{\pi}}$, the maximizer of $L_R(\boldsymbol{\pi})$. To improve numerical stabilities, we exclude the haplotypes whose estimated frequencies are 0 or very close to 0, i.e. less than $\max(2/n, 0.001)$. The MLE of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ is consistent, asymptotically normal, and asymptotically efficient.

Note that the likelihood for case-control studies was previously given in Lin et al. [2008] and is reformulated in this section to conform with the notation for the imputation method. The likelihood for cross-sectional studies is new.

RESULTS

SIMULATION STUDIES

We carried out extensive simulation studies to evaluate the performance of the MLE and imputation methods in realistic settings. We generated genotype data for various sets of five SNPs according to the LD patterns observed in the HapMap CEU sample. For each SNP set, we chose one SNP to be untyped in the study data. For some SNP sets, we picked more than one SNP to be untyped, one at a time, each representing a different scenario. Table I lists the nine scenarios used in the simulation studies, with R^2 [Stram, 2004] ranging from 0.41 to 0.98.

We explored three types of association: (1) single-SNP effects, (2) gene-environment interactions, and (3) multi-SNP effects. For each type of model, we considered both cross-sectional and case-control designs. Since the case-control design naturally requires a binary trait, we focused on quantitative traits for cross-sectional studies. Thus, there were six series of simulation studies. For each setup, we simulated 10,000 data sets with 2,000 study subjects and 60 trios. Under the case-control design, we set the overall disease rate to be approximately 1% and selected an equal number of cases and controls. We chose 60 trios for the reference panel so as to approximate the CEU sample in the current version (i.e. phase 3) of the HapMap database, which consists of 44 trios, 8 duos, and 17 singletons. For each simulated data set, we applied the MLE and imputation approaches. For the latter approach, we imputed the unknown genotype by both the dosage and the most likely genotype, which are referred to as the IMP-DOS and IMP-MLG methods, respectively. All the analysis was based on the Wald statistic.

Our first series of simulation studies was concerned with the (marginal) effect of an untyped SNP on a quantitative trait in a cross-sectional study. We generated the trait value from the linear regression model

$$Y=\alpha+\beta G_u+\varepsilon,$$

where ε is standard normal and α = 0. Table II displays the results for various values of β. As expected, the MLE is virtually unbiased in all cases. IMP-DOS also shows negligible bias, which is not surprising because conditional mean imputation is known to yield consistent estimators of regression parameters under the linear model [Little, 1992]. The estimator of β produced by IMP-MLG is seriously biased toward zero and the bias can be as much as 25% of the true parameter value. For non-zero β, both IMP-DOS and IMP-MLG tend to underestimate the variances, so their confidence intervals have poor coverage probabilities. Under scenario S8, in which $R^2 = 0.98$, the coverage probability of the 99% confidence interval of IMP-DOS is only 98% when β = 0.9. IMP-MLG is much worse than IMP-DOS because it suffers from both biased estimation of parameter and underestimation of variance; see S1–S3. As predicted by our theory, both IMP-DOS and IMP-MLG have appropriate type I error. In some cases (i.e. S2, S3, and S5), IMP-DOS is slightly more powerful than MLE. This phenomenon is attributed to the underestimation of variance by IMP-DOS. When R^2 is large (e.g. S7–S9), all methods have the same power.

In our second series of studies, we simulated case-control data under the logistic regression model

$$\Pr(Y=1|G_u)=\frac{e^{\alpha+\beta G_u}}{1+e^{\alpha+\beta G_u}},$$

where α was set to −4.6 to yield disease rates of approximately 1%. The results are summarized in Table III. Unlike linear regression, IMP-DOS can produce substantial bias under logistic regression; see S1–S3 and S5. MLE is now uniformly more powerful than both IMP-DOS and IMP-MLG; this feature can be seen more clearly in Figure 1. The power gain of MLE over imputation persists as R^2 approaches 1 because MLE exploits the HWE assumption, whereas imputation does not. When R^2 is low, the bias of imputation (under non-linear models) also affects its power. Again, all three methods have accurate control of type I error. As in cross-sectional studies, both IMP-DOS and IMP-MLG tend to underestimate the variances (for non-zero β) and thus yield poor confidence interval coverage, especially when β is large and R^2 is low.

Our third and fourth series of studies were focused on gene-environment interactions under the cross-sectional and case-control designs, respectively. We generated data from the same models as in the first two series but with the linear predictors $\alpha+\beta_1 G_u+\beta_2 X+\beta_3 G_u X$, where X is Bernoulli with $\Pr(X = 1) = 0.4$. The results for cross-sectional studies are displayed in Table IV. For detecting interactions, both IMP-DOS and IMP-MLG produce confidence intervals with poor coverage probabilities, especially when the effects are large and the LD is low; see S1–S6. Both may lose control of type I error and are substantially less powerful than MLE. The power gain of MLE is largely attributed to its incorporation of gene-environment independence. The power difference decreases as R^2 increases. In the extreme case of $R^2 = 1$, the summation in (4) disappears and MLE is equivalent to imputation. The results for case-control studies are shown in Table V. Both imputation methods yield biased estimates, poor confidence interval coverage, and diminished power. The power difference between MLE and imputation is further illustrated in Figure 2. The power gain of MLE is again largely attributed to its use of gene-environment independence. If we analyzed the imputed genotypes (either the dosage or the most likely genotype) by the method of Chatterjee and Carroll [2005], which also exploits gene-environment independence, then the power gain of MLE was reduced considerably (results not shown).

Our last two series of studies dealt with multi-SNP effects. We set the untyped SNP to be causal and included all five SNPs in the joint analysis. For making inference on the effect of the untyped SNP, the performance of IMP-DOS and IMP-MLG is similar to the first two series of studies (results not shown). In particular, type I error is properly controlled. This is not surprising because our theory indicates that imputation yields a valid test of the untyped SNP even when there are environmental factors or typed SNPs in the model. On the other hand, if the untyped SNP is associated with the trait, the bias in the estimation of its effect can cause bias in estimating the null effects of the typed SNPs. Indeed, both IMP-DOS and IMP-MLG can have inflated type I error in testing the effects of the typed SNPs and the inflation of type I error becomes more severe as the effect of the untyped SNP increases. Figures 3 and 4 display these results for cross-sectional and case-control studies, respectively. As before, MLE has accurate control of type I error.

WTCCC DATA

We considered WTCCC data on type 1 diabetes (T1D). The database contains 1,963 subjects with T1D and 2,938 controls. For the typed SNPs, we applied the standard

Armitage trend test. For the untyped SNPs that are cataloged in the HapMap phase 3 database, we applied both MLE and IMP-DOS, with the phase 3 HapMap CEU sample as the reference panel. For each untyped SNP, we first identified the typed SNPs within 100 kb and then found a set of four that yields the largest R^2. If there were fewer than eight SNPs within 100 kb, we enlarged the window until a minimum of eight SNPs were located. If there were more than 20 SNPs within 100 kb, we restricted our attention to the closest 20 SNPs so as to reduce computation time.

As shown in Figure 5, MLE and IMP-DOS produce nearly identical quantile-quantile (Q-Q) plots for the untyped SNPs, which are similar to that of the typed SNPs. The deviations of the test statistics from the null distribution are minor except in the extreme tails, which correspond to significant associations. The over-dispersion parameter (i.e. the genomic control λ) was estimated at approximately 1.05 for all three plots. These results illustrate that, for single-SNP analysis, both MLE and imputation have correct type I error.

Figure 6 displays the results of the association tests for both typed and untyped SNPs on chromosomes 1, 6, and 12, which have the strongest evidence of association. Both MLE and IMP-DOS were able to identify untyped SNPs that are more strongly associated with the disease than typed SNPs, but MLE picked out those SNPs more clearly. This is not surprising since MLE is expected to be more powerful than imputation.

DISCUSSION

We have presented two approaches to the analysis of untyped SNPs and investigated their properties both theoretically and numerically. The maximum-likelihood approach yields approximately unbiased parameter estimators, proper confidence intervals, and accurate control of type I error. It tends to be more powerful than the imputation approach, especially for case-control studies and in testing gene-environment interactions. The maximum-likelihood method requires the study sample and reference panel be generated from the same underlying population and may be numerically unstable when the haplotype frequencies are low.

We have assumed gene-environment independence in the maximum-likelihood approach. This assumption is satisfied in most applications and can substantially improve the efficiency of association analysis, especially in case-control studies [Chatterjee and Carroll, 2005]. It is possible to allow gene-environment dependence, but the analysis will be more complicated and less efficient.

The imputation approach has some advantages over the maximum-likelihood approach. Numerically, the former is more stable than the latter. For single-SNP tests, imputation has proper control of type I error even if the reference panel does not match the study population. For testing other hypotheses, however, imputation may have inflated type I error. In general, imputation yields biased parameter estimators and incorrect variance estimators. Because the bias can be upward and the variance is underestimated, imputation can sometimes be more powerful than maximum likelihood. Thus, maximum likelihood and imputation are complementary to each other. One possible strategy is to use imputation (with the dosage as the imputed value) in the initial single-SNP tests and to use maximum likelihood for more complex analysis once a region of disease association has been identified.

For cross-sectional studies, Xie and Stram [2005] showed that the score test based on the dosage of the risk haplotype is asymptotically valid. We have shown that imputation is asymptotically valid for single-SNP tests under both cross-sectional and case-control designs whether the untyped SNP is imputed by the dosage or the most likely genotype.

Note that the haplotype analysis does not involve external data, whereas the analysis of untyped SNPs does.

Because it ignores the random variation of the reference panel, the imputation approach generally underestimates the variances of the parameter estimators. As the size of the reference panel increases, the underestimation of variance becomes less severe and thus confidence intervals have better coverage probabilities. The size of the reference panel, however, has little influence on the bias of imputation. On the other hand, increasing the size of the reference panel reduces the variance of the MLE. Indeed, the power of maximum likelihood improves at a faster rate than imputation as the reference panel becomes larger, especially under the case-control design (results not shown).

Both MLE and imputation are computationally fast, and the relevant software is available at our website. It took about 8 hr on a 64-bit, 30-GHz Intel Xeon machine (Chapel Hill, NC) to perform the MLE analysis on chromosome 1 of the WTCCC GWAS data. Imputation was slightly faster. The computational savings of imputation will be more substantial if there are multiple traits of interest because the untyped SNPs only need to be imputed once.

For computational expediency, we used the significance level of 0.01 in our simulation studies. The relatively small number of replicates required for obtaining accurate summary statistics at this significance level allowed us to explore a very wide variety of scenarios. It would be formidable to conduct extensive simulation studies at the significance level of 10^{-4} or lower, which would require at least 1 million replicates. We repeated some of our simulation studies using the significance level of 10^{-4}, and the basic conclusions regarding the relative merits of MLE and imputation remained the same.

We have focused on tagging-based imputation. An alternative approach is to use hidden-Markov models (HMM) [Browning and Browning, 2007; Marchini et al., 2007; Li et al., 2008, submitted]. The latter approach, which explores the LD information over a larger region and incorporates population genetics knowledge, can yield more accurate prediction of untyped genotypes in certain situations. We chose tagging over HMM in this article for several reasons: (1) using the same amount of information to infer missing genotypes ensures that the maximum likelihood and imputation methods are compared on equal footing; (2) an investigation by the imputation working group of GAIN [The GAIN Collaborative Research Group, 2007] revealed that tagging is nearly as accurate as HMM (unpublished data); and (3) tagging is much simpler and faster than HMM and can handle much larger studies. We are currently trying to incorporate HMM into the maximum-likelihood framework. The conclusions of this article regarding the relative merits of the maximum likelihood versus imputation approaches are expected to hold when tagging is replaced by HMM.

Acknowledgments

Contract grant sponsor: National Institutes of Health.

This research was supported by the National Institutes of Health and a Gillings Innovation Laboratory (GIL) award at the UNC Gillings School of Global Public Health. The authors thank the reviewers for their helpful comments.

References

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81:1084–1097. [PubMed: 17924348]

Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika. 2005; 92:399–418.

de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005; 37:1217–1223. [PubMed: 16244653]

Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). J Am Stat Assoc. 2006; 101:89–118.

Lin DY, Hu Y, Huang BE. Simple and efficient analysis of disease association with missing genotype data. Am J Hum Genet. 2008; 82:444–452. [PubMed: 18252224]

Little RJA. Regression with missing X's: a review. J Am Stat Assoc. 1992; 87:1227–1237.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39:906–913. [PubMed: 17572673]

Nicolae DL. Testing untyped alleles (TUNA)—applications to genome-wide association studies. Genet Epidemiol. 2006; 30:718–727. [PubMed: 16986160]

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–441.

Stram D. Tag SNP selection for association studies. Genet Epidemiol. 2004; 27:365–374. [PubMed: 15372618]

The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: The Genetic Association Information Network. Nat Genet. 2007; 39:1045–1051. [PubMed: 17728769]

The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]

Xie R, Stram D. Asymptotic equivalence between two score tests for haplotype-specific risk in general linear models. Genet Epidemiol. 2005; 29:166–170. [PubMed: 16025443]

Zaitlen N, Kang HM, Eskin E, Halperin E. Leveraging the HapMap correlation structure in association studies. Am J Hum Genet. 2007; 80:683–691. [PubMed: 17357074]

APPENDIX A: THEORETICAL PROPERTIES OF IMPUTATION

We are interested in the effect of the untyped SNP genotype G_u on the phenotype Y adjusted for the effects of covariates $\mathbf{W}$ (if there are any). The covariates, which are required to be fully observed, may include environmental factors and typed SNPs and are allowed to be correlated with the untyped SNP. The linear predictor is assumed to take the form of $\alpha+\beta_{G_u}G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}$, where β_{G_u} and $\beta_{\mathbf{W}}$ represent the regression effects of G_u and $\mathbf{W}$, respectively. Write $\boldsymbol{\beta}=(\beta_{G_u},\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}})^{\mathrm{T}}$. We are particularly interested in testing the null hypothesis H_0: $\beta_{G_u}=0$.

Let n denotes the total number of study subjects. For $i = 1, \ldots, n$, let Y_i, G_{ui}, and $\mathbf{W}_i$ denote the values of Y, G_u, and $\mathbf{W}$ on the ith subject. We replace G_{ui} by $\hat{G}_{ui}$, where $\hat{G}_{ui}$ is the imputed value of G_{ui} based on Equation (3), and then apply standard likelihood methods to the imputed data set $(Y_i, \hat{G}_{ui}, \mathbf{W}_i)$ $(i = 1, \ldots, n)$. The validity of such analysis does not follow from the standard likelihood theory because the n imputed values $\{\hat{G}_{ui}\}(i = 1, \ldots, n)$ are correlated due to the presence of the estimator $\tilde{\pi}$ in them.

We first consider cross-sectional studies. The "likelihood" for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$ based on the imputed data set takes the form $L(\boldsymbol{\theta})=\prod_{i=1}^{n}P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\widehat{G}_{ui},\mathbf{W}_i)$. Denote the resulting estimator by $\hat{\boldsymbol{\theta}}$. As mentioned above, standard likelihood theory is not applicable to $\hat{\boldsymbol{\theta}}$ because the n terms in $L(\theta)$ are not independent.

Under H_0: $\beta_{G_u} = 0$, Y is related to $\mathbf{W}$ only and is independent of G_u given $\mathbf{W}$. Assume that G_t is independent of Y given $\mathbf{W}$. (This assumption holds if G_t is independent of Y or is part of $\mathbf{W}$.) Then $\hat{G}_u$, which is a function of G_t and $\tilde{\boldsymbol{\pi}}$, is also independent of Y given $\mathbf{W}$, regardless of the value of $\tilde{\boldsymbol{\pi}}$. In other words, the regression effects of $\hat{G}_u$ and $\mathbf{W}$ on Y are the same as those of G_u and $\mathbf{W}$ under H_0. Denote the reference panel by R. Conditional on R, the imputed values are uncorrelated, so that, under H_0, the random vector $\mathbf{I}^{1/2}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a multivariate normal distribution with mean zero and identity covariance matrix, where $\mathbf{I}(\boldsymbol{\theta}) = -\partial^2 \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$. Because the limiting distribution does not depend on R, the convergence also holds unconditionally. Thus, standard likelihood methods can be used to test H_0 (even if the study sample and reference panel are drawn from different populations).

The above result hinges critically on the null hypothesis H_0: $\beta_{G_u} = 0$ under the linear predictor $\alpha+\beta_{G_u} G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}$, which ensures that $\hat{\boldsymbol{\theta}}$ converges to the true value of $\boldsymbol{\theta}$ conditional on R. If $\beta_{G_u} \neq 0$, then the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ conditional on R depends on R, so that the inverse information matrix $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$, which ignores the variability in the reference panel, will underestimate the true variation of $\hat{\boldsymbol{\theta}}$. Thus, the confidence intervals for β_{G_u} will not have proper coverage probabilities unless $\beta_{G_u} = 0$. It should also be pointed out that the validity of association testing is not guaranteed if the linear predictor does not take the form of $\alpha+\beta_{G_u} G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}$.

We now consider the analysis of case-control data under the logistic regression model

$$\Pr(Y=1|G_u, \mathbf{W})=\frac{\mathrm{e}^{\alpha+\beta_{G_u} G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}}}{1+\mathrm{e}^{\alpha+\beta_{G_u} G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}}}.$$

Write $\boldsymbol{\theta}=(\alpha,\beta_{G_u},\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}})^{\mathrm{T}}$. If G_u were observed on all study subjects, then the maximum-likelihood estimator of $\boldsymbol{\theta}$ (based on the prospective likelihood) would converge to $\boldsymbol{\theta}^*$ and its covariance matrix would be consistently estimated by the inverse information matrix, where $\boldsymbol{\theta}^*$ is the same as $\boldsymbol{\theta}$ except that α is replaced by a different constant [Prentice and Pyke, 1979]. Let $\hat{\boldsymbol{\theta}}$ be the maximizer of the (prospective) likelihood based on the imputed data:

$$L(\boldsymbol{\theta})=\prod_{i=1}^{n}\frac{\mathrm{e}^{Y_i(\alpha+\beta_{G_u}\widehat{G}_{ui}+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}_i)}}{1+\mathrm{e}^{\alpha+\beta_{G_u}\widehat{G}_{ui}+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}_i}}.$$

It then follows from the above arguments for cross-sectional studies that, under H_0: $\beta_{G_u} = 0$, the random vector $\mathbf{I}^{1/2}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges to a multivariate normal distribution with mean zero and identity covariance matrix, where $\mathbf{I}(\boldsymbol{\theta}) = -\partial^2\log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$. Thus, the association testing is valid. Again, the variance is underestimated by the inverse information matrix if $\beta_{G_u} \neq 0$, and the association testing may not be valid for other types of hypotheses.

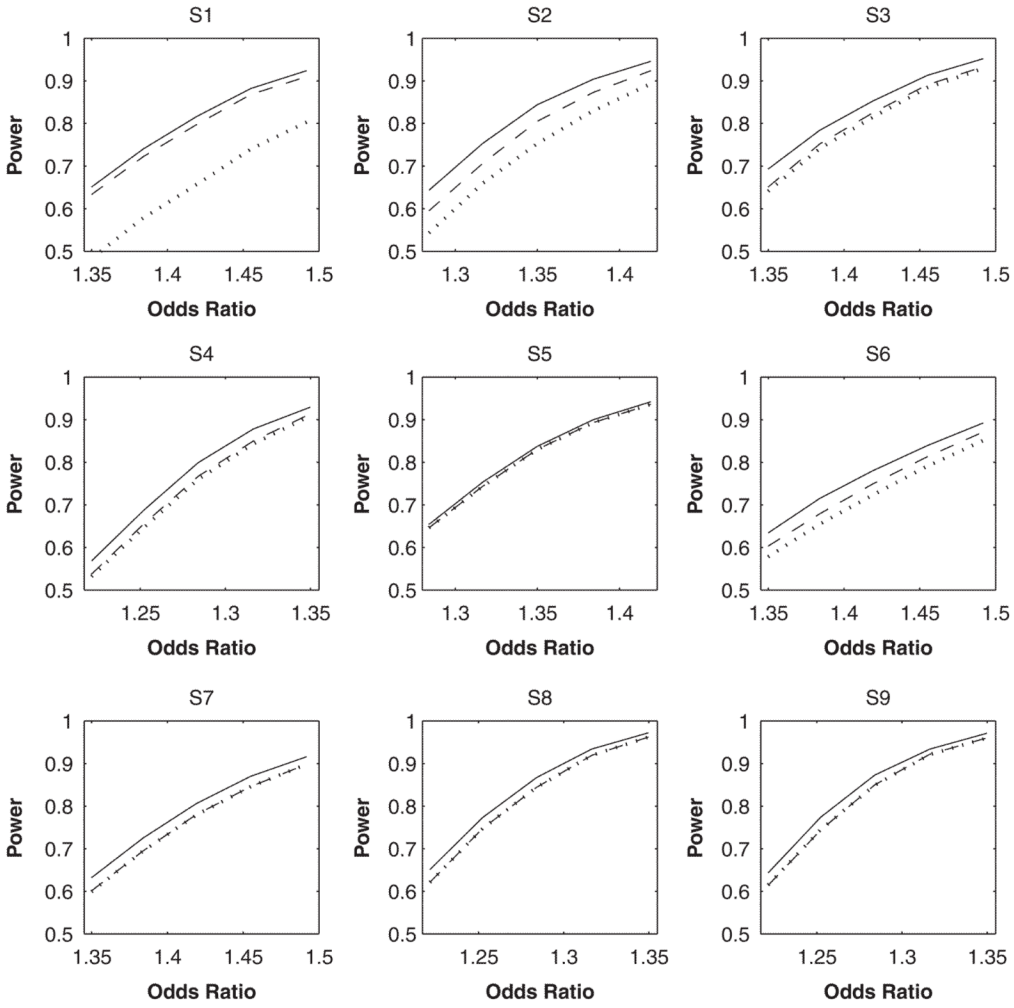

Fig. 1.

Power of testing the effect of an untyped SNP at the 1% nominal significance level under the case-control design. The solid, dashed, and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively. SNP, single nucleotide polymorphism; MLE, maximum-likelihood estimator.

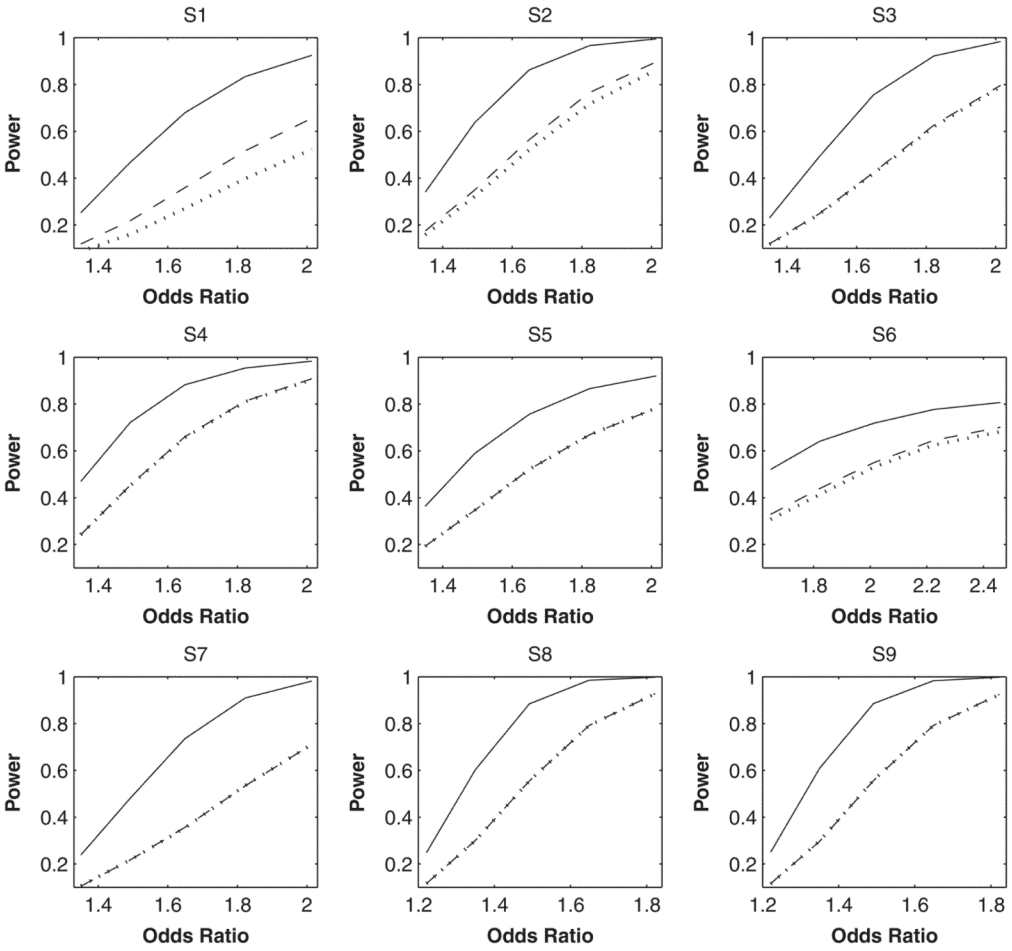

Fig. 2.

Power of testing gene-environment interactions at the 1% nominal significance level under the case-control design. The solid, dashed, and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively. MLE, maximum-likelihood estimator.

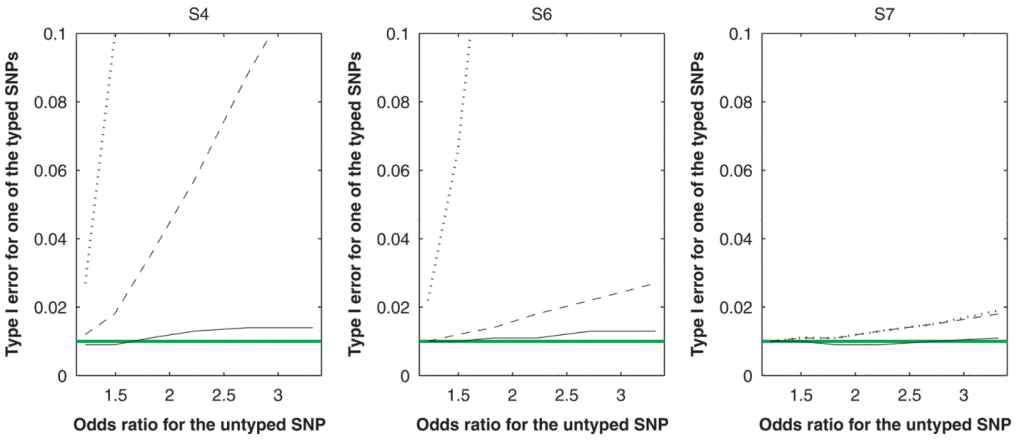

Fig. 3.

Type I error for testing the null effect of a typed SNP on a quantitative trait at the 1% nominal significance level in the joint analysis involving a causal, untyped SNP under the cross-sectional design. The solid, dashed, and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively. SNP, single nucleotide polymorphism; MLE, maximum-likelihood estimator.

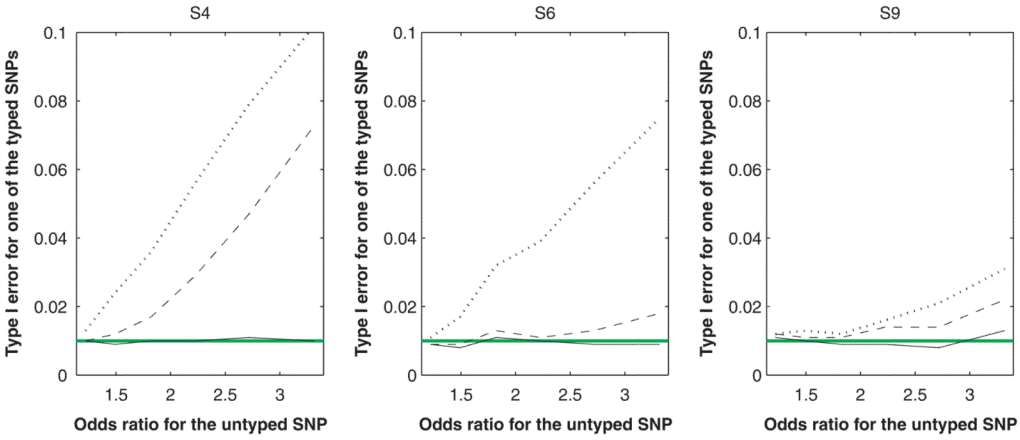

Fig. 4.

Type I error for testing the null effect of a typed SNP at the 1% nominal significance level in the joint analysis involving a causal, untyped SNP under the case-control design. The solid, dashed, and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively. SNP, single nucleotide polymorphism; MLE, maximum-likelihood estimator.

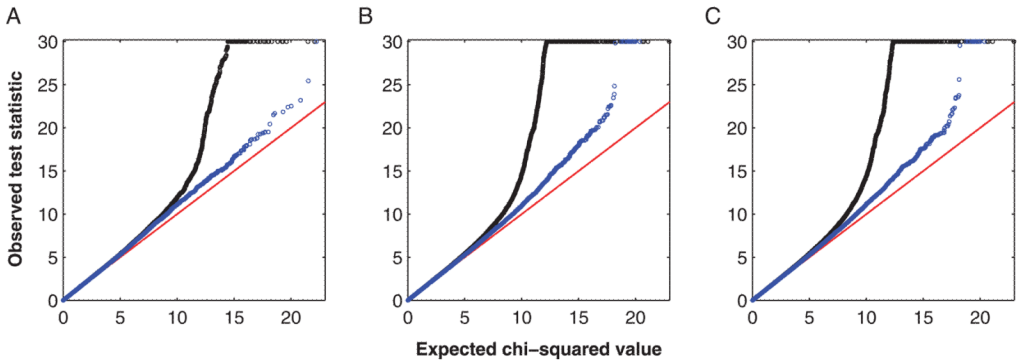

Fig. 5.

Q-Q plots for the single SNP analysis of the T1D data from the WTCCC study: (A) Armitage trend test for typed SNPs, (B) MLE for untyped SNPs, and (C) IMP-DOS for untyped SNPs. Chi-squared statistics exceeding 30 are truncated. The black curve in (A) pertains to 392,746 typed SNPs that pass the standard project filters, have MAF >1% and missing data rates <1%, and have good cluster plots. The black curves in (B) and (C) pertain to 819,727 untyped SNPs that are cataloged in Phase 3 of HapMap with MAF >1%. The Q-Q plots, which exclude all SNPs located in the regions of association listed in Table III of the WTCCC [2007] paper, are superimposed in blue. The blue curves show that departures in the extreme tails of the distributions of test statistics are due to regions with strong signals for association. WTCCC, Wellcome Trust Case-Control Consortium; SNP, single nucleotide polymorphism; T1D, type 1 diabetes; MAF, minor allele frequencies; MLE, maximum-likelihood estimator; Q-Q, quantile-quantile.

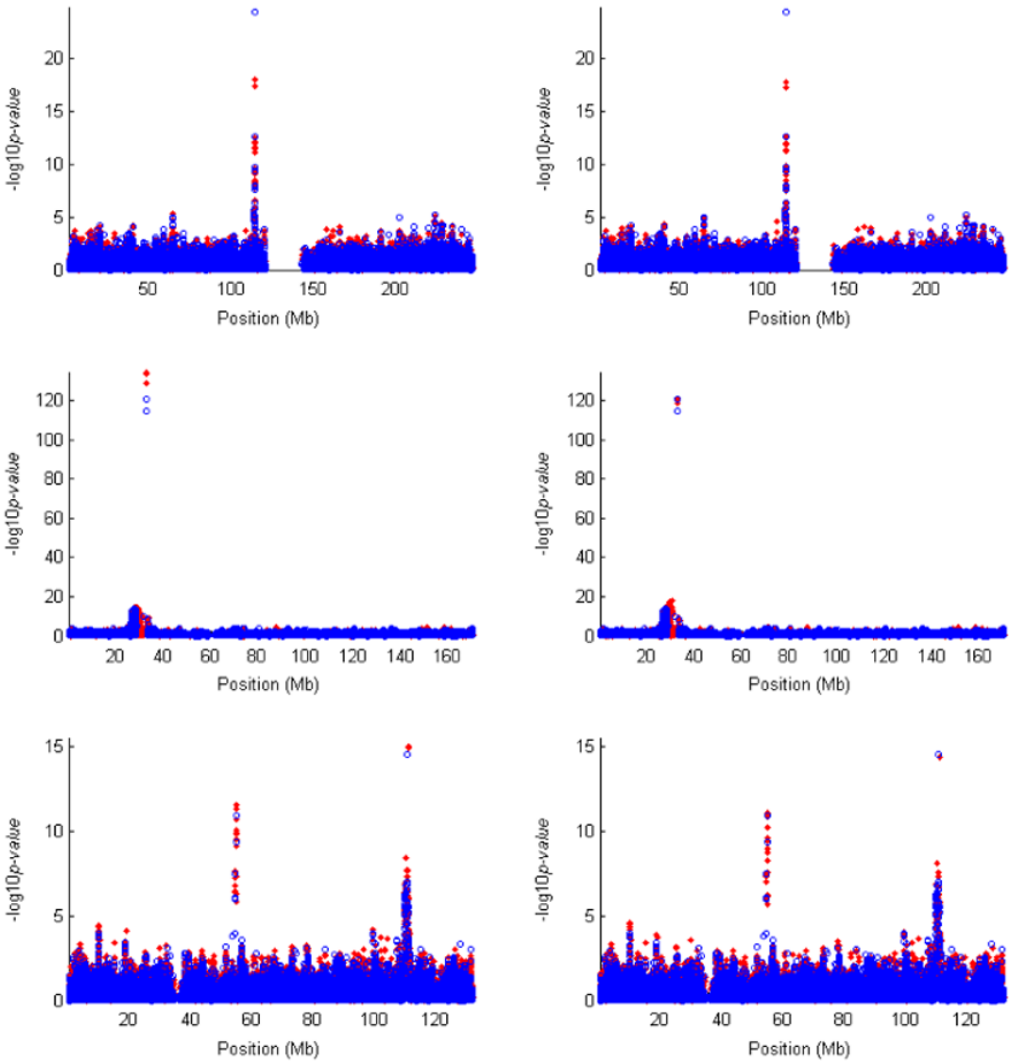

Fig. 6.

Results of single-SNP association tests for the WTCCC study of T1D. The $\log_{10}$ *P*-values for typed SNPs and untyped SNPs are shown in blue circles and red dots, respectively. The three rows correspond to chromosomes 1, 6 and 12, which have the strongest evidence of association. The left column corresponds to the trend test for the typed SNPs and the MLE method for the untyped SNPs. The right column corresponds to the trend test for the typed SNPs and the IMP-DOS method for the untyped SNPs. All typed SNPs pass the standard project filters, have MAF >1% and missing data rate <1%, and have good cluster plots. All untyped SNPs have MAF >1% in HapMap. WTCCC, Wellcome Trust Case-Control Consortium; SNP, single nucleotide polymorphism; T1D, type 1 diabetes; MLE, maximum-likelihood estimator.

TABLE I

Haplotype frequencies for the scenarios used in simulation studies

	S1: R^2 = 0.41, MAF = 0.39		**S2: R^2 = 0.59, MAF = 0.28**		**S3: R^2 = 0.70, MAF = 0.15**	
Haplotype	**UTTTT**	**Frequency**	**TTTUT**	**Frequency**	**TTUTT**	**Frequency**
h_1	00011	0.0513	00100	0.3171	00000	0.0513
h_2	00100	0.0260	00101	0.0988	00100	0.1460
h_3	01011	0.0855	00111	0.2027	01000	0.6958
h_4	01100	0.3094	01001	0.1518	10011	0.1069
h_5	01101	0.1377	10100	0.1059		
h_6	11011	0.0085	10101	0.0209		
h_7	11100	0.1775	10111	0.0793		
h_8	11101	0.0247	11001	0.0235		
h_9	11111	0.1794				
	S4: R^2 = 0.81, MAF = 0.33		S5: R^2 = 0.84, MAF = 0.24		S6: R^2 = 0.93, MAF = 0.15	
Haplotype	UTTTT	Frequency	TTUTT	Frequency	TTUTT	Frequency
h_1	00011	0.2846	01101	0.2852	00011	0.0513
h_2	00101	0.0128	10100	0.2510	00100	0.0260
h_3	00111	0.0342	11000	0.2393	01011	0.0855
h_4	10101	0.2374	11100	0.0321	01100	0.3094
h_5	10111	0.1917	11101	0.0963	01101	0.1377
h_6	11110	0.2393	11110	0.0961	11011	0.0085
h_7					11100	0.1775
h_8					11101	0.0247
h_9					11111	0.1794
	S7: R^2 = 0.95, MAF = 0.09		S8: R^2 = 0.98, MAF = 0.28		S9: R^2 = 0.98, MAF = 0.29	
Haplotype	UTTTT	Frequency	TTUTT	Frequency	TTTTU	Frequency
h_1	00111	0.3809	01000	0.4231	01000	0.4231
h_2	01110	0.2350	01010	0.1154	01010	0.1154
h_3	01111	0.2900	01011	0.0043	01011	0.0043

	S1: R^2 = 0.41, MAF = 0.39		S2: R^2 = 0.59, MAF = 0.28		S3: R^2 = 0.70, MAF = 0.15	
Haplotype	**UTTTT**	**Frequency**	**TTTUT**	**Frequency**	**TTUTT**	**Frequency**
h_4	11001	0.0897	01111	0.2821	01111	0.2821
h_5	11111	0.0044	10010	0.1751	10010	0.1751

"U" and "T" indicate the untyped and typed SNP positions, respectively. R^2 is the squared correlation between the expected and true allele counts [Stram, 2004]. MAF is the minor allele frequency of the untyped SNP.

TABLE II

Simulation results for studying the effect of an untyped SNP on a quantitative trait under the cross-sectional design

		MLE					IMP-DOS					IMP-MLG				
	β	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	0.0	0.000	0.051	0.051	0.989	0.011	0.000	0.051	0.051	0.990	0.010	0.000	0.041	0.041	0.989	0.011
	0.1	0.000	0.051	0.051	0.990	0.274	−0.001	0.051	0.051	0.990	0.267	−0.031	0.041	0.041	0.965	0.190
	0.2	0.000	0.052	0.051	0.990	0.905	−0.002	0.051	0.051	0.990	0.903	−0.062	0.042	0.041	0.857	0.781
	0.6	0.000	0.053	0.052	0.990	1.00	−0.007	0.056	0.053	0.987	1.00	−0.185	0.047	0.043	0.063	1.00
	0.9	−0.002	0.054	0.052	0.988	1.00	−0.010	0.061	0.056	0.982	1.00	−0.277	0.053	0.046	0.001	1.00
S2	0.0	0.000	0.046	0.046	0.991	0.009	0.000	0.046	0.046	0.990	0.010	0.000	0.036	0.036	0.991	0.009
	0.1	0.000	0.046	0.047	0.991	0.325	0.000	0.046	0.046	0.990	0.336	−0.026	0.036	0.036	0.965	0.309
	0.2	0.000	0.048	0.048	0.992	0.960	0.000	0.048	0.046	0.988	0.963	−0.051	0.038	0.036	0.843	0.937
	0.6	0.001	0.057	0.057	0.988	1.00	−0.001	0.061	0.047	0.954	1.00	−0.154	0.048	0.037	0.151	1.00
	0.9	0.000	0.060	0.060	0.985	1.00	−0.001	0.077	0.049	0.903	1.00	−0.231	0.060	0.038	0.059	1.00
S3	0.0	0.000	0.055	0.054	0.992	0.008	0.000	0.055	0.054	0.991	0.009	0.000	0.041	0.041	0.990	0.010
	0.1	0.000	0.055	0.055	0.992	0.217	0.001	0.055	0.054	0.989	0.237	−0.025	0.041	0.041	0.977	0.233
	0.2	0.001	0.057	0.057	0.991	0.856	0.001	0.058	0.054	0.986	0.871	−0.050	0.042	0.041	0.907	0.863
	0.6	0.001	0.070	0.070	0.987	1.00	0.004	0.078	0.055	0.936	1.00	−0.149	0.046	0.041	0.172	1.00
	0.9	0.000	0.074	0.073	0.987	1.00	0.006	0.100	0.056	0.863	1.00	−0.224	0.051	0.042	0.033	1.00
S4	0.0	0.000	0.037	0.037	0.989	0.011	0.000	0.037	0.037	0.989	0.011	0.000	0.035	0.035	0.989	0.011
	0.1	0.000	0.037	0.037	0.989	0.544	0.000	0.037	0.037	0.989	0.544	−0.007	0.035	0.035	0.986	0.540
	0.2	0.000	0.038	0.037	0.989	0.998	−0.001	0.037	0.037	0.988	0.998	−0.013	0.035	0.035	0.983	0.997
	0.6	0.000	0.038	0.038	0.988	1.00	−0.002	0.039	0.038	0.986	1.00	−0.039	0.036	0.036	0.929	1.00
	0.9	0.000	0.039	0.038	0.989	1.00	−0.003	0.041	0.038	0.984	1.00	−0.059	0.036	0.036	0.829	1.00
S5	0.0	0.000	0.041	0.040	0.991	0.009	0.000	0.041	0.040	0.991	0.009	0.000	0.036	0.036	0.991	0.009
	0.1	0.000	0.041	0.041	0.991	0.456	0.000	0.041	0.040	0.990	0.463	−0.012	0.036	0.036	0.985	0.463
	0.2	0.001	0.042	0.042	0.991	0.991	0.001	0.042	0.040	0.988	0.991	−0.023	0.036	0.036	0.971	0.991
	0.6	0.001	0.048	0.048	0.988	1.00	0.001	0.050	0.041	0.966	1.00	−0.071	0.036	0.036	0.722	1.00
	0.9	0.001	0.052	0.051	0.987	1.00	0.002	0.060	0.041	0.929	1.00	−0.106	0.037	0.036	0.365	1.00
S6	0.0	0.000	0.050	0.050	0.990	0.010	0.000	0.050	0.050	0.990	0.010	0.000	0.048	0.047	0.990	0.010
	0.1	0.000	0.050	0.050	0.990	0.286	0.000	0.050	0.050	0.990	0.284	−0.008	0.048	0.047	0.988	0.266
	0.2	0.000	0.050	0.050	0.990	0.919	−0.001	0.050	0.050	0.990	0.918	−0.015	0.048	0.047	0.985	0.904

Genet Epidemiol. Author manuscript; available in PMC 2011 December 1.

		MLE					IMP-DOS					IMP-MLG				
	β	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
	0.6	0.000	0.050	0.050	0.990	1.00	−0.003	0.050	0.050	0.990	1.00	−0.046	0.048	0.048	0.943	1.00
	0.9	0.000	0.050	0.049	0.989	1.00	−0.004	0.051	0.051	0.991	1.00	−0.069	0.049	0.049	0.876	1.00
S7	0.0	0.000	0.056	0.056	0.990	0.010	0.000	0.056	0.056	0.990	0.010	0.000	0.055	0.055	0.990	0.010
	0.1	0.000	0.056	0.056	0.990	0.215	0.000	0.056	0.056	0.990	0.214	0.000	0.055	0.055	0.990	0.215
	0.2	0.000	0.056	0.056	0.990	0.844	0.000	0.056	0.056	0.990	0.844	−0.001	0.055	0.055	0.990	0.845
	0.6	0.000	0.056	0.056	0.990	1.00	0.000	0.056	0.056	0.990	1.00	−0.003	0.055	0.055	0.990	1.00
	0.9	0.000	0.056	0.056	0.990	1.00	0.000	0.056	0.056	0.990	1.00	−0.004	0.056	0.056	0.990	1.00
S8	0.0	0.000	0.036	0.036	0.990	0.010	0.000	0.036	0.036	0.990	0.010	0.000	0.035	0.035	0.990	0.010
	0.1	0.000	0.036	0.036	0.990	0.590	0.000	0.036	0.036	0.990	0.590	−0.002	0.035	0.035	0.991	0.590
	0.2	0.000	0.036	0.036	0.991	0.999	0.000	0.036	0.036	0.990	0.999	−0.003	0.035	0.035	0.991	0.999
	0.6	0.000	0.037	0.037	0.990	1.00	0.000	0.037	0.036	0.987	1.00	−0.009	0.035	0.035	0.988	1.00
	0.9	0.000	0.038	0.038	0.989	1.00	0.000	0.038	0.036	0.982	1.00	−0.014	0.035	0.035	0.984	1.00
S9	0.0	0.000	0.036	0.035	0.990	0.010	0.000	0.036	0.035	0.990	0.010	0.000	0.035	0.035	0.990	0.010
	0.1	0.000	0.036	0.035	0.990	0.599	0.000	0.036	0.035	0.990	0.599	−0.001	0.035	0.035	0.990	0.598
	0.2	0.000	0.036	0.035	0.990	0.999	0.000	0.036	0.035	0.990	0.999	−0.001	0.035	0.035	0.990	0.999
	0.6	0.000	0.036	0.035	0.990	1.00	0.000	0.036	0.035	0.990	1.00	−0.004	0.035	0.035	0.991	1.00
	0.9	−0.001	0.036	0.035	0.990	1.00	−0.001	0.036	0.035	0.989	1.00	−0.005	0.035	0.035	0.990	1.00

S1–S9 denote the nine scenarios listed in Table I. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. PW is the type I error/power for testing zero parameter value at the 0.01 nominal significance level. Each entry is based on 10,000 replicates.

TABLE III

Simulation results for studying the effect of an untyped SNP on the risk of disease under the case-control design

		MLE					IMP-DOS					IMP-MLG				
	β	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	0.0	0.001	0.100	0.099	0.989	0.011	0.001	0.101	0.101	0.990	0.010	0.001	0.082	0.082	0.990	0.010
	0.3	0.000	0.103	0.101	0.990	0.651	−0.010	0.101	0.100	0.989	0.633	−0.093	0.082	0.081	0.922	0.483
	0.6	0.000	0.108	0.106	0.989	0.999	−0.031	0.102	0.099	0.984	0.999	−0.190	0.082	0.081	0.580	0.993
	0.9	−0.008	0.115	0.113	0.986	1.00	−0.070	0.104	0.100	0.961	1.00	−0.297	0.083	0.082	0.156	1.00
S2	0.0	−0.002	0.091	0.090	0.991	0.009	−0.001	0.093	0.092	0.989	0.011	−0.001	0.073	0.072	0.988	0.012
	0.3	0.000	0.085	0.084	0.989	0.844	0.012	0.094	0.091	0.987	0.805	−0.067	0.075	0.072	0.934	0.753
	0.6	−0.002	0.084	0.082	0.990	1.00	0.046	0.103	0.092	0.971	1.00	−0.117	0.083	0.073	0.757	1.00
	0.9	−0.009	0.084	0.083	0.989	1.00	0.093	0.115	0.094	0.919	1.00	−0.157	0.097	0.075	0.566	1.00
S3	0.0	−0.002	0.106	0.106	0.992	0.008	−0.001	0.109	0.108	0.989	0.011	0.000	0.082	0.082	0.989	0.011
	0.3	−0.001	0.099	0.098	0.991	0.693	0.012	0.110	0.105	0.987	0.651	−0.067	0.081	0.079	0.953	0.641
	0.6	0.000	0.096	0.096	0.990	1.00	0.047	0.120	0.103	0.969	1.00	−0.116	0.082	0.078	0.841	1.00
	0.9	−0.003	0.096	0.097	0.989	1.00	0.096	0.136	0.103	0.915	1.00	−0.153	0.087	0.078	0.682	1.00
S4	0.0	0.000	0.073	0.073	0.990	0.010	0.000	0.075	0.074	0.989	0.011	0.000	0.071	0.070	0.989	0.011
	0.3	0.001	0.075	0.074	0.990	0.929	0.002	0.078	0.077	0.990	0.913	−0.016	0.074	0.073	0.986	0.909
	0.6	0.000	0.077	0.077	0.989	1.00	0.007	0.082	0.081	0.990	1.00	−0.028	0.077	0.077	0.984	1.00
	0.9	−0.006	0.082	0.081	0.990	1.00	0.011	0.088	0.087	0.989	1.00	−0.039	0.083	0.082	0.979	1.00
S5	0.0	0.000	0.079	0.079	0.989	0.011	0.000	0.082	0.081	0.989	0.011	0.000	0.072	0.071	0.989	0.011
	0.3	0.001	0.086	0.087	0.990	0.837	−0.005	0.085	0.084	0.988	0.830	−0.040	0.074	0.074	0.977	0.830
	0.6	0.002	0.103	0.102	0.989	1.00	−0.023	0.093	0.088	0.980	1.00	−0.092	0.078	0.077	0.912	1.00
	0.9	0.009	0.130	0.127	0.988	1.00	−0.051	0.102	0.093	0.961	1.00	−0.154	0.082	0.081	0.744	1.00
S6	0.0	0.000	0.096	0.097	0.991	0.009	0.000	0.100	0.100	0.991	0.009	0.000	0.095	0.095	0.991	0.009
	0.3	0.001	0.102	0.103	0.990	0.634	0.000	0.105	0.106	0.989	0.603	−0.022	0.100	0.101	0.989	0.578
	0.6	−0.002	0.112	0.112	0.989	0.999	−0.007	0.114	0.114	0.990	0.997	−0.049	0.108	0.108	0.981	0.996
	0.9	−0.004	0.122	0.121	0.989	1.00	−0.014	0.125	0.123	0.987	1.00	−0.077	0.118	0.117	0.964	1.00
S7	0.0	0.000	0.109	0.108	0.991	0.009	0.000	0.112	0.111	0.992	0.008	0.000	0.112	0.111	0.992	0.008
	0.3	−0.001	0.103	0.102	0.989	0.632	−0.001	0.106	0.106	0.991	0.601	−0.002	0.106	0.105	0.991	0.599
	0.6	0.000	0.097	0.098	0.990	1.00	0.001	0.102	0.102	0.990	1.00	−0.002	0.101	0.101	0.990	1.00
	0.9	−0.003	0.096	0.095	0.989	1.00	−0.001	0.101	0.100	0.989	1.00	−0.005	0.100	0.099	0.989	1.00

		MLE					IMP-DOS					IMP-MLG				
	β	**Bias**	**SE**	**SEE**	**CP**	**PW**	**Bias**	**SE**	**SEE**	**CP**	**PW**	**Bias**	**SE**	**SEE**	**CP**	**PW**
S8	0.0	0.000	0.070	0.069	0.989	0.011	0.000	0.072	0.071	0.989	0.011	0.000	0.071	0.070	0.989	0.011
	0.3	0.000	0.066	0.067	0.990	0.972	0.002	0.069	0.070	0.990	0.962	−0.003	0.068	0.069	0.989	0.962
	0.6	−0.001	0.066	0.066	0.990	1.00	0.004	0.071	0.070	0.990	1.00	−0.006	0.069	0.069	0.989	1.00
	0.9	−0.006	0.066	0.066	0.990	1.00	0.005	0.072	0.071	0.990	1.00	−0.008	0.070	0.070	0.991	1.00
S9	0.0	0.000	0.069	0.069	0.988	0.012	0.000	0.071	0.071	0.989	0.011	0.000	0.071	0.070	0.989	0.011
	0.3	0.000	0.067	0.067	0.992	0.971	0.000	0.069	0.069	0.991	0.960	−0.001	0.069	0.069	0.991	0.960
	0.6	−0.002	0.066	0.066	0.991	1.00	−0.001	0.069	0.069	0.992	1.00	−0.004	0.069	0.069	0.992	1.00
	0.9	−0.007	0.067	0.066	0.989	1.00	−0.002	0.072	0.071	0.989	1.00	−0.007	0.071	0.070	0.988	1.00

S1–S9 denote the nine scenarios listed in Table I. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. PW is the type I error/power for testing zero parameter value at the 0.01 nominal significance level. Each entry is based on 10,000 replicates.

TABLE IV

Simulation results for studying gene-environment interactions under the cross-sectional design with a quantitative trait

			MLE					IMP-DOS					IMP-MLG				
	β_1	β_3	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	0.5	0.0	0.000	0.097	0.096	0.990	0.010	−0.002	0.107	0.107	0.991	0.009	−0.001	0.087	0.087	0.990	0.010
	0.5	0.2	0.000	0.095	0.094	0.990	0.328	−0.004	0.108	0.109	0.991	0.210	−0.063	0.088	0.089	0.970	0.150
	0.5	0.3	0.000	0.094	0.093	0.990	0.740	−0.005	0.108	0.111	0.992	0.531	−0.094	0.089	0.090	0.943	0.385
	0.5	0.9	0.000	0.092	0.091	0.990	1.00	−0.012	0.115	0.121	0.993	1.00	−0.279	0.098	0.100	0.409	1.00
	1.2	0.0	0.000	0.085	0.084	0.991	0.009	−0.002	0.116	0.123	0.994	0.006	−0.001	0.099	0.102	0.992	0.008
S2	0.5	0.0	−0.001	0.092	0.092	0.991	0.009	−0.002	0.097	0.096	0.990	0.010	−0.001	0.076	0.075	0.991	0.009
	0.5	0.2	−0.001	0.092	0.091	0.990	0.352	−0.002	0.098	0.097	0.989	0.303	−0.053	0.077	0.076	0.964	0.273
	0.5	0.3	−0.001	0.092	0.091	0.990	0.763	−0.002	0.100	0.098	0.989	0.682	−0.078	0.078	0.076	0.926	0.633
	0.5	0.9	−0.001	0.091	0.090	0.991	1.00	−0.003	0.118	0.103	0.977	1.00	−0.232	0.092	0.081	0.377	1.00
	1.2	0.0	0.000	0.087	0.087	0.989	0.011	−0.002	0.107	0.105	0.989	0.011	−0.001	0.083	0.083	0.992	0.009
S3	0.5	0.0	0.000	0.110	0.110	0.991	0.009	0.000	0.113	0.112	0.989	0.011	0.000	0.085	0.084	0.989	0.011
	0.5	0.2	0.001	0.110	0.110	0.990	0.224	0.001	0.115	0.112	0.987	0.223	−0.050	0.086	0.085	0.974	0.218
	0.5	0.3	0.000	0.111	0.111	0.989	0.560	0.002	0.118	0.113	0.986	0.541	−0.075	0.087	0.085	0.951	0.533
	0.5	0.9	0.000	0.110	0.111	0.991	1.00	0.006	0.144	0.116	0.964	1.00	−0.224	0.094	0.087	0.487	1.00
	1.2	0.0	0.000	0.108	0.108	0.990	0.010	−0.001	0.124	0.116	0.984	0.017	0.000	0.093	0.088	0.984	0.016
S4	0.5	0.0	−0.001	0.077	0.076	0.988	0.012	−0.001	0.077	0.077	0.990	0.010	−0.001	0.073	0.072	0.989	0.011
	0.5	0.2	−0.001	0.077	0.076	0.988	0.516	−0.002	0.077	0.077	0.990	0.496	−0.014	0.073	0.073	0.988	0.488
	0.5	0.3	−0.001	0.077	0.076	0.988	0.913	−0.002	0.078	0.077	0.991	0.901	−0.021	0.073	0.073	0.987	0.895
	0.5	0.9	−0.001	0.076	0.075	0.988	1.00	−0.004	0.079	0.080	0.990	1.00	−0.060	0.074	0.076	0.966	1.00
	1.2	0.0	−0.001	0.076	0.075	0.988	0.012	−0.001	0.080	0.080	0.990	0.010	−0.001	0.076	0.076	0.990	0.010
S5	0.5	0.0	−0.001	0.084	0.083	0.990	0.011	−0.001	0.084	0.083	0.989	0.011	−0.001	0.074	0.073	0.989	0.011
	0.5	0.2	−0.001	0.084	0.083	0.989	0.432	−0.001	0.085	0.083	0.988	0.432	−0.025	0.075	0.073	0.985	0.432
	0.5	0.3	−0.001	0.085	0.084	0.989	0.842	−0.001	0.086	0.084	0.988	0.840	−0.036	0.075	0.074	0.978	0.840
	0.5	0.9	−0.001	0.085	0.084	0.989	1.00	0.000	0.097	0.085	0.978	1.00	−0.108	0.077	0.075	0.872	1.00
	1.2	0.0	−0.001	0.084	0.083	0.988	0.012	−0.001	0.089	0.086	0.987	0.013	−0.001	0.078	0.076	0.987	0.013
S6	0.5	0.0	−0.001	0.102	0.101	0.989	0.011	0.000	0.102	0.102	0.990	0.010	−0.001	0.097	0.098	0.990	0.010
	0.5	0.2	−0.001	0.102	0.101	0.989	0.277	−0.001	0.102	0.103	0.990	0.261	−0.016	0.098	0.098	0.990	0.245
	0.5	0.3	−0.001	0.101	0.101	0.989	0.649	−0.002	0.102	0.103	0.990	0.621	−0.024	0.098	0.098	0.989	0.591

	β_1	β_3	MLE					IMP-DOS					IMP-MLG				
			Bias	**SE**	**SEE**	**CP**	**PW**	**Bias**	**SE**	**SEE**	**CP**	**PW**	**Bias**	**SE**	**SEE**	**CP**	**PW**
	0.5	0.9	−0.001	0.101	0.100	0.989	1.00	−0.005	0.103	0.105	0.991	1.00	−0.070	0.099	0.100	0.973	1.00
	1.2	0.0	−0.001	0.099	0.099	0.989	0.011	0.000	0.103	0.105	0.991	0.009	−0.001	0.100	0.101	0.991	0.009
S7	0.5	0.0	−0.002	0.114	0.114	0.991	0.009	−0.002	0.114	0.114	0.991	0.009	−0.002	0.114	0.113	0.991	0.009
	0.5	0.2	−0.002	0.114	0.114	0.991	0.202	−0.002	0.114	0.114	0.991	0.202	−0.003	0.114	0.113	0.991	0.201
	0.5	0.3	−0.002	0.114	0.114	0.990	0.515	−0.002	0.114	0.114	0.991	0.512	−0.004	0.114	0.113	0.991	0.512
	0.5	0.9	−0.004	0.114	0.114	0.990	1.00	−0.003	0.114	0.114	0.991	1.00	−0.007	0.114	0.114	0.991	1.00
	1.2	0.0	−0.002	0.114	0.114	0.990	0.010	−0.002	0.115	0.114	0.991	0.009	−0.002	0.114	0.114	0.991	0.009
S8	0.5	0.0	0.001	0.073	0.073	0.989	0.011	0.001	0.074	0.073	0.989	0.011	0.001	0.072	0.072	0.989	0.011
	0.5	0.2	0.001	0.074	0.073	0.989	0.572	0.001	0.074	0.073	0.989	0.572	−0.002	0.072	0.072	0.989	0.572
	0.5	0.3	0.001	0.074	0.073	0.989	0.939	0.001	0.074	0.073	0.990	0.939	−0.004	0.072	0.072	0.989	0.939
	0.5	0.9	−0.001	0.074	0.073	0.989	1.00	0.001	0.075	0.073	0.989	1.00	−0.013	0.073	0.072	0.987	1.00
	1.2	0.0	0.001	0.074	0.073	0.990	0.011	0.001	0.074	0.073	0.989	0.011	0.001	0.073	0.072	0.989	0.011
S9	0.5	0.0	0.001	0.073	0.072	0.990	0.010	0.001	0.073	0.072	0.990	0.010	0.001	0.073	0.072	0.990	0.010
	0.5	0.2	0.001	0.073	0.072	0.990	0.579	0.001	0.073	0.072	0.990	0.577	0.000	0.073	0.072	0.990	0.577
	0.5	0.3	0.000	0.073	0.072	0.990	0.943	0.001	0.073	0.072	0.990	0.942	−0.001	0.073	0.072	0.990	0.943
	0.5	0.9	−0.001	0.073	0.072	0.990	1.00	0.000	0.073	0.073	0.990	1.00	−0.005	0.073	0.072	0.989	1.00
	1.2	0.0	0.001	0.073	0.072	0.990	0.010	0.001	0.073	0.073	0.990	0.011	0.001	0.073	0.072	0.990	0.010

$\beta_2 = 0.2$. S1–S9 denote the nine scenarios listed in Table I. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. PW is the type I error/power for testing zero parameter value at the 0.01 nominal significance level. Each entry is based on 10,000 replicates.

TABLE V

Simulation results for studying gene-environment interactions under the case-control design

		MLE					IMP-DOS					IMP-MLG				
	β_3	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	0.0	0.000	0.151	0.149	0.990	0.010	−0.002	0.210	0.209	0.990	0.010	−0.002	0.169	0.169	0.991	0.009
	0.5	−0.008	0.165	0.163	0.988	0.679	−0.021	0.218	0.216	0.988	0.359	−0.157	0.175	0.175	0.952	0.270
	0.9	−0.025	0.188	0.187	0.987	0.985	−0.065	0.233	0.231	0.985	0.859	−0.295	0.187	0.187	0.840	0.751
S2	0.0	0.000	0.136	0.135	0.992	0.008	−0.001	0.189	0.189	0.991	0.009	−0.001	0.148	0.147	0.991	0.009
	0.5	−0.006	0.140	0.139	0.991	0.863	0.032	0.198	0.194	0.989	0.565	−0.103	0.157	0.152	0.965	0.521
	0.9	−0.023	0.153	0.153	0.989	1.00	0.098	0.216	0.204	0.978	0.990	−0.153	0.174	0.160	0.927	0.980
S3	0.0	0.002	0.162	0.160	0.992	0.008	0.001	0.226	0.223	0.991	0.009	0.000	0.170	0.168	0.990	0.010
	0.5	−0.002	0.160	0.158	0.990	0.756	0.034	0.232	0.223	0.987	0.424	−0.100	0.171	0.168	0.973	0.421
	0.9	−0.016	0.167	0.166	0.987	1.00	0.099	0.247	0.227	0.977	0.968	−0.152	0.177	0.171	0.948	0.967
S4	0.0	−0.002	0.110	0.109	0.988	0.012	−0.002	0.154	0.153	0.990	0.010	−0.002	0.145	0.144	0.990	0.010
	0.5	−0.011	0.130	0.129	0.989	0.882	0.004	0.169	0.169	0.990	0.660	−0.026	0.160	0.159	0.989	0.657
	0.9	−0.036	0.157	0.158	0.990	0.996	0.007	0.193	0.194	0.990	0.977	−0.044	0.182	0.183	0.989	0.977
S5	0.0	−0.001	0.118	0.118	0.992	0.008	0.001	0.166	0.166	0.991	0.009	0.001	0.146	0.146	0.991	0.009
	0.5	−0.011	0.150	0.150	0.991	0.756	−0.014	0.187	0.185	0.989	0.521	−0.072	0.163	0.163	0.982	0.521
	0.9	−0.031	0.200	0.199	0.991	0.963	−0.056	0.221	0.216	0.987	0.904	−0.157	0.191	0.190	0.965	0.904
S6	0.0	−0.003	0.146	0.146	0.990	0.010	−0.003	0.205	0.206	0.992	0.008	−0.003	0.196	0.195	0.993	0.007
	0.5	−0.017	0.186	0.185	0.991	0.521	−0.005	0.238	0.235	0.990	0.329	−0.041	0.226	0.223	0.988	0.307
	0.9	−0.049	0.243	0.241	0.993	0.806	−0.023	0.282	0.280	0.991	0.701	−0.087	0.268	0.266	0.989	0.682
S7	0.0	0.003	0.165	0.163	0.991	0.009	0.000	0.230	0.230	0.990	0.010	0.000	0.229	0.229	0.990	0.010
	0.5	−0.004	0.157	0.156	0.991	0.735	−0.001	0.227	0.226	0.988	0.356	−0.003	0.226	0.225	0.989	0.354
	0.9	−0.015	0.157	0.156	0.990	1.00	0.000	0.225	0.226	0.992	0.926	−0.004	0.224	0.225	0.991	0.926
S8	0.0	0.001	0.104	0.104	0.991	0.009	0.000	0.146	0.146	0.991	0.009	0.000	0.144	0.144	0.991	0.009
	0.5	−0.007	0.107	0.107	0.990	0.985	0.005	0.149	0.149	0.991	0.792	−0.003	0.147	0.147	0.991	0.792
	0.9	−0.025	0.116	0.116	0.986	1.00	0.008	0.159	0.157	0.989	0.999	−0.006	0.156	0.155	0.989	0.999
S9	0.0	0.001	0.104	0.103	0.991	0.009	0.000	0.146	0.145	0.990	0.010	0.000	0.145	0.145	0.990	0.010
	0.5	−0.007	0.108	0.107	0.989	0.983	0.002	0.148	0.149	0.989	0.792	−0.001	0.147	0.148	0.990	0.792
	0.9	−0.026	0.117	0.116	0.986	1.00	0.000	0.158	0.157	0.991	0.999	−0.006	0.157	0.156	0.990	0.999

$\beta_1 = 0.0$, $\beta_2 = 0.1$. S1–S9 denote the nine scenarios listed in Table I. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. PW is the type I error/power for testing zero parameter value at the 0.01 nominal significance level. Each entry is based on 10,000 replicates.