

Evaluating Haplotype Effects in Case-Control
Studies via Penalized-Likelihood Approaches:
Prospective or Retrospective Analysis?

Megan L. Koehler¹

Howard D. Bondell¹

Jung-Ying Tzeng^{1,2*}

¹Department of Statistics, North Carolina State University, Raleigh, North Carolina

²Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina

*Correspondence to: Jung-Ying Tzeng, Department of Statistics, Campus Box 7566, North
Carolina State University, Raleigh, NC 27695.

Email: jytzeng@stat.ncsu.edu

Tel: 919-513-2723

Fax: 919-515-7315

SUMMARY

Penalized likelihood methods have become increasingly popular in recent years for evaluating haplotype-phenotype association in case-control studies. Although a retrospective likelihood is dictated by the sampling scheme, these penalized methods are typically built upon prospective likelihoods due to their modeling simplicity and computational feasibility. It has been well documented that for unpenalized methods, prospective analyses of case-control data can be valid but less efficient than their retrospective counterparts when testing for association, and result in substantial bias when estimating the haplotype effects. For penalized methods, which combine effect estimation and testing in one step, the impact of using a prospective likelihood is not clear. In this work, we examine the consequences of ignoring the sampling scheme for haplotype-based penalized likelihood methods. Our results suggest that the impact of prospective analyses depends on (1) the underlying genetic mode and (2) the genetic model adopted in the analysis. When the correct genetic model is used, the difference between the two analyses is negligible for additive and slight for dominant haplotype effects. For recessive haplotype effects, the more appropriate retrospective likelihood clearly outperforms the prospective likelihood. If an additive model is incorrectly used, as the true underlying genetic is unknown a priori, both retrospective and prospective penalized methods suffer from a sizeable power loss and increase in bias. The impact of using the incorrect genetic model is much bigger on retrospective analyses than prospective analyses, and results in comparable performances for both methods.

Key Words: haplotype-based association analysis; variable selection; regularized regression; prospective likelihood; retrospective likelihood

INTRODUCTION

Haplotype-based association analysis evaluates the joint effects of closely linked genetic markers on a trait of interest. When compared to its single-marker counterparts, this multi-marker approach can be more powerful to detect associations when the causal variants are not genotyped [de Bakker et al., 2005; Zaitlen et al., 2007], have low frequency [de Bakker et al., 2005; Schaid, 2004], or exhibit cis-acting effects [Clark, 2004; Schaid, 2004]. A standard approach for performing haplotype-based analysis is to regress the trait value on the haplotypes and test the significance of the regression parameters [Balding, 2006]. In recent years, applying penalized likelihood methods to identify important haplotypic factors has become increasingly popular in the literature. For example, Li et al. [2007] use the least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996] to perform selection among numerous possible haplotypes resulting from different haplotype window lengths. Guo and Lin [2009] use LASSO regression to evaluate the effects of rare haplotypes and high-dimension haplotype–environment interactions. Tzeng et al. [2010] use adaptive LASSO regression [Zou, 2006] to study high dimensional gene-treatment interactions in a haplotype-based pharmacogenetic analysis. These methods introduce a penalty on the regression coefficients and shrink the coefficient estimates of non-important covariates towards zero. The motivation behind using penalized methods in haplotype-based analysis is that while the predictor space under consideration may be large, many of the haplotypic predictors are not likely to be associated with the phenotype. In this case, it is more efficient to shrink these effect estimates to zero than to estimate them purely. This shrinkage leads to a reduction in variance and can increase the power to detect important haplotypic predictors [Guo and Lin, 2009].

Modifications of classic penalized methods have also been developed to perform haplotype-based analysis and attempt to address issues specific to this type of analysis. Tanck and colleagues [Souverein et al., 2006, 2008; Tanck et al., 2003] use a modified version of Ridge regression to stabilize inference for rare haplotypes. By constructing an L_2 -norm penalty term on the differences in coefficients of similar haplotypes, the coefficients of rare haplotypes are smoothed towards that of a similar common haplotype. Tzeng and Bondell [2010] modify traditional adaptive LASSO regression by placing an L_1 -norm penalty on pair-wise differences of the regression coefficients. This allows for effect comparisons between all pairs of distinct haplotypes, rather than with respect to an arbitrary baseline haplotype, during the estimation process. As result, the approach is able to sort haplotypes into different groups according to their effect sizes and eliminates the need for a post-hoc pair-wise analysis of haplotype effects. The key of a penalized regression method lies in the form of the penalty – by carefully designing the form of the penalty, one can gear the penalized-likelihood approach towards accomplishing various desired tasks.

Penalized regression methods rely on the underlying data likelihood. When analyzing data from case-control studies, one can implement methods based on a prospective likelihood (modeling the probability of disease status conditional on exposure) or a retrospective likelihood (modeling the probability of exposure conditional on disease status). Under a case-control design, a retrospective likelihood should be used because data are collected based on disease status. However, in practice, it is common for researchers to use a prospective likelihood, as it does not require specifying a model for the joint distribution of the genetic and environmental effects. Bypassing this step makes implementing prospective methods much easier than retrospective methods [Lin et al, 2005]. This approach seems congruent with the well-known

result that optimizing the prospective likelihood yields the same inference on the disease model parameters as optimizing the retrospective likelihood [Prentice and Pyke, 1979]. This result requires that the distribution of the covariates be free of restrictions, which does not generally hold in haplotype-based analysis. Haplotypes are not directly observed from unphased genotype data. In order to reconstruct the haplotypes and estimate their effects, some assumptions must be placed on their frequency distribution (typically Hardy-Weinberg equilibrium).

Most of the penalized regression approaches mentioned above utilized a prospective likelihood. It has been well-documented that when using non-penalized regression methods in haplotype-based analysis of case-control data, ignoring the ascertainment scheme can be detrimental. A prospective analysis can lead to a loss of efficiency and severe bias when assessing the haplotype effects [Cordell, 2006; Satten and Epstein, 2004; Stram et al., 2003]. The aim of this work is to determine whether similar consequences occur when using penalized regression for case-control studies. Specifically, we consider the adaptive LASSO penalty, and examine the relative performance in parameter estimation and model selection between the penalized method using a prospective likelihood and using a retrospective likelihood. In subsequent sections, we describe the methods used to address the question posed by this work. We illustrate the use of these methods through extensive simulation studies and close by discussing the results.

METHODS

PROSPECTIVE AND RETROSPECTIVE LIKELIHOODS

Let the vector (Y_i, G_i, E_i) represent the observed data for individual i in a case-control sample of size n . Let Y_i be a binary indicator of disease status where $Y_i = 1$ if individual i is a

case and 0 otherwise. Let G_i denote the unphased genotype of individual i at m biallelic SNPs and E_i denote any environmental covariates measured on individual i . Let H_i represent the vector of haplotype counts for individual i . Although researchers want to investigate the relationship between Y_i and H_i , they only have access to G_i and the individual's haplotype set must be inferred from their unphased genotypes.

The relationship between the disease phenotype and the covariates can be characterized by the conditional density function $P(Y|H, E)$. A standard approach for binary trait values is logistic regression which models the conditional probability as

$$P(Y = y|H, E) = \frac{\exp\{y \cdot (\beta_0 + \mathcal{Z}(H, E)^T \beta)\}}{1 + \exp\{\beta_0 + \mathcal{Z}(H, E)^T \beta\}}$$

where β_0 is an intercept, β is the vector of disease model parameters representing the log-odds ratios, and $\mathcal{Z}(H, E)$ is a specified vector-valued function of H and E .

Various likelihood models have been developed to conduct inference about the disease model parameters in haplotype-based analyses while properly accounting for phase uncertainty. The inference can be based on a prospective likelihood or on a retrospective likelihood. In this work, we consider maximum likelihood methods developed by two groups – one focusing on a prospective approach and the other on a retrospective approach. We implement the prospective method developed in Lake et al [2003]. Their prospective likelihood models $P(Y_i|G_i, E_i)$ and is expressed as

$$L_P = \prod_{i=1}^n P(Y_i|G_i, E_i) = \prod_{i=1}^n \sum_{H_i \in S(G_i)} P(H_i, Y_i|G_i, E_i) = \prod_{i=1}^n \sum_{H_i \in S(G_i)} P(Y_i|H_i, E_i)P(H_i).$$

where $S(G)$ is the set of all haplotype pairs consistent with G , $P(H = h) = 2 \prod_{j=1}^l \pi_j^{h_j} / h_j$ under the assumption of Hardy-Weinberg Equilibrium, h_k is the number of copies of the k^{th} haplotype

in H , π_k is the population frequency of the k^{th} haplotype, and l is the number of haplotypes included in the disease model. We implement the retrospective method developed in Lin and Zeng [2006]. Their retrospective likelihood models $P(G_i, E_i|Y_i)$ and is expressed as

$$L_R = \prod_{i=1}^n P(G_i, E_i|Y_i) = \prod_{i=1}^n \sum_{H_i \in \overline{S}(G_i)} P(H_i, G_i, E_i|Y_i) \propto \prod_{i=1}^n \sum_{H_i \in \overline{S}(G_i)} P(Y_i|H_i, E_i)P(H_i)P(E_i|G_i).$$

The only difference between the two likelihoods is the conditional density function $P(E_i|G_i)$ found in the retrospective likelihood. The parameters in this model are of no interest to researchers performing haplotype-based association analysis, but they must be estimated in order to make proper inference when using a retrospective design. Specifying a model for this conditional density function and the subsequent maximum likelihood estimation are computationally intensive. As a result, researchers often rely on prospective methods when analyzing case-control data even though retrospective methods are dictated by the ascertainment scheme [Lin and Zeng, 2006].

HAPLOTYPE ANALYSIS VIA PENALIZED LIKELIHOOD METHODS

While many different penalized likelihood methods can be used in haplotype-based association analysis, we consider the adaptive LASSO (ALASSO) penalty in this work. This approach achieves simultaneous variable selection and parameter estimation and is an oracle procedure. This refers to the fact that the approach asymptotically selects the correct model, and the resulting estimator is root-n consistent and asymptotically normal with the same variance as if the true model were known before hand [Zou, 2006].

The ALASSO effect estimates are obtained by minimizing a penalized negative log-likelihood. These estimates are expressed as

$$\hat{\beta}_\lambda = \operatorname{argmin}_\beta -\ell_n(\beta, \phi) + \lambda \sum_{j=1}^p w_j |\beta_j|,$$

where $\ell_n(\beta, \phi)$ denotes the log-likelihood, ϕ is a (possible) set of nuisance parameters (e.g. the haplotype frequencies, π_k), λ is the non-negative regularization parameter that controls the amount of shrinkage, and w_j are data-dependent weights. By placing an L_1 -norm penalty on the regression coefficients, the ALASSO can set their estimates to exactly zero if the value of λ is large enough. It is this feature that allows the procedure to perform simultaneous variable selection and parameter estimation. Unlike its predecessor the LASSO, the ALASSO places a different penalty on each coefficient through the use of adaptive weights that are inversely proportional to their relative importance. Consequently, haplotypes with negligible effects receive larger penalties and are more readily shrunk to zero. This allows the effects of associated haplotypes to be estimated more efficiently. Zou [2006] proposed to set the weights as $w_j = |\tilde{\beta}_j|^{-\gamma}$, where $\tilde{\beta}_j$ is an initial root-n consistent estimator of β_j and $\gamma > 0$ is an additional tuning parameter. In our analysis, we chose $\gamma = 1$ and let $\tilde{\beta}_j$ be the maximum likelihood estimate of the haplotype effect computed by `haplo.glm` in R and HAPSTAT in Linux for the prospective and retrospective likelihoods, respectively [Lake et al, 2003; Lin et al, 2005]. When using penalized regression, the design matrix should be scaled so that the penalization is applied equally across all predictors. Because we chose $\gamma = 1$ and $\tilde{\beta}_j$ are scale equivariant estimators (i.e., a scale change in the design matrix changes the regression coefficients by a power of the same scale), the ALASSO automatically controls for scaling differences in the adaptive weights, bypassing the need to scale the imputed haplotype design matrix.

The ALASSO solution ($\hat{\beta}_\lambda$) also depends on the value of λ . The regularization parameter controls the tradeoff between model fit and model sparsity. By including more predictors, one

can continually improve the fit on the training data at the expense of interpretability and over fitting. Many model selection criteria, like Mallorw’s C_p , Akaike information criterion (AIC), Bayesian information criterion (BIC), and cross validation, can be used to determine the appropriate value of λ from an exhaustive grid search. Because the goal of haplotype-based association analysis is more aligned with selecting the true model than minimizing prediction error, we use BIC for tuning which can achieve consistent model selection [Yang, 2005]. BIC is defined as

$$BIC = -2\ell_n(\hat{\beta}_\lambda, \hat{\phi}) + df_\lambda \cdot \log(\frac{1}{n})$$

where $\ell_n(\hat{\beta}_\lambda, \hat{\phi})$ is the log-likelihood evaluated at the estimated regression coefficients and maximized over ϕ for a given λ and df_λ is the degrees of freedom, which equals the number of non-zero elements in $(\hat{\beta}_\lambda, \hat{\phi})$. The λ that minimizes BIC is chosen as the regularization parameter, and its corresponding $\hat{\beta}_\lambda$ is the ALASSO estimate.

For computational convenience, the least squares approximation (LSA) method was used to calculate the ALASSO solution. The LSA method replaces the objective function of the original ALASSO problem with its asymptotically equivalent least squares form [Weng and Lang, 2007]. The method is motivated by a standard Taylor series expansion of $-\ell_n(\beta, \phi)$ about $(\tilde{\beta}, \tilde{\phi})$, the function’s unpenalized minimizer, and shows that the ALASSO estimate can asymptotically be given as

$$\hat{\beta}_\lambda = argmin_\beta (\beta - \tilde{\beta})^T \tilde{\Sigma}^{-1} (\beta - \tilde{\beta}) + \lambda \sum_{j=1}^p w_j |\beta_j|$$

where $\tilde{\Sigma}$ is the estimated covariance matrix of $\tilde{\beta}$. Because the underlying data likelihoods are not quadratic in the regression coefficients, transforming them into their asymptotically equivalent forms greatly reduces the computational costs for finding the ALASSO solution [Weng and

Lang, 2007]. Using the LSA method eliminates the need for an iterative procedure to perform optimization; it only requires one unpenalized fit of the original objective function and then a grid search to determine λ . The final estimate is again chosen by minimizing the BIC.

SIMULATION STUDIES

We performed simulation studies to examine the performance of the ALASSO method under two competing data likelihoods when analyzing case-control data. Specifically, we wanted to determine if using a prospective likelihood in place of the more appropriate retrospective likelihood was detrimental when performing haplotype-based analyses using a penalized likelihood method. To answer this question, we compared the parameter estimation and model selection properties of each approach. For ease of discussion, let *aPro* refer to ALASSO coupled with a prospective likelihood and *aRetro* refer to ALASSO coupled with a retrospective likelihood.

SIMULATION SETTINGS

Our simulation studies were based on two haplotype distributions (given in Table 1) studied by Lin and Huang [2007]. These distributions are based on the common haplotypes formed by five SNPs on chromosome 18 in the CEU sample of the HapMap data. The SNPs used to build the first haplotype distribution were in strong linkage disequilibrium, while those used to build the second haplotype distribution were not. Distribution 1 represents a haplotype distribution with a few high frequency haplotypes, while the haplotype frequencies in Distribution 2 are more uniform. Each distribution was normalized so that the haplotype frequencies summed to 1.

For each haplotype distribution, we considered two simulation studies – one in which a single haplotype was associated with the disease (Simulation I) and one in which two haplotypes were associated with the disease (Simulation II). Because our focus was on identifying and estimating disease-haplotype associations, only genetic covariates were considered in our simulation studies (i.e. E is taken to be \emptyset). In both simulation studies, we examined the effect of varying the genetic mode of the associated haplotype(s) on the performance of aPro and aRetro. We took the sample size to be $n = 1,000$ with an equal number of cases and controls. We allowed the associated haplotype(s) to act additively, dominantly, or recessively with respect to disease risk. In practice, the genetic mode of a risk haplotype is unknown a priori, and researchers typically analyze the data additively regardless of the true genetic model. To mimic this scenario, we analyzed each dataset using the correct genetic model and again using an additive model.

In addition to genetic mode, we varied the frequency and effect size of the associated haplotype. In Simulation I, a rare or a common haplotype was chosen to be the associated genetic variant for each haplotype distribution. A haplotype with frequency less than 0.10 was considered rare; otherwise it was considered common. We set the effect sizes of the associated haplotype (in terms of the odds ratio θ) so that the power of finding the effect fell in a reasonable range. For additive and dominant models, we set $\theta = \{1.0, 1.3, 1.5, 1.7, 2.0\}$, and for recessive models, we set $\theta = \{1.0, 2.0, 2.5, 3.0, 3.5\}$. We let $\theta = 1$ to examine the performance of the approaches under a null model. In Simulation II, we allowed two haplotypes to be associated with the disease, where the associated haplotypes were both rare, one rare and one common, or both common. The odds ratios of both associated haplotypes were set to $\theta = 1.7$ for additive and dominant models and $\theta = 3.0$ for recessive models. The settings for Simulation I and Simulation II can be found in Table 2. In all, 78 different simulation settings were studied.

DATA GENERATION

We generated the haplotype pair of an individual conditional on their disease status and then dissolved the haplotype pair into its unphased genotypes. Let $P(H = h|Y = y)$ denote the probability of having a particular haplotype pair conditional on disease status. This probability can be expressed as

$$P(H = h|Y = y) = \frac{P(Y = y|H = h) \cdot P(H = h)}{\sum_h P(Y = y|H = h) \cdot P(H = h)}.$$

For a case individual, $P(Y = 1|H = h)$ was found using the logistic regression model

$$P(Y = 1|H) = \frac{\exp\{\beta_0 + \mathcal{Z}(H)^T \beta\}}{1 + \exp\{\beta_0 + \mathcal{Z}(H)^T \beta\}}.$$

For a control individual, $P(Y = 0|H = h) = 1 - P(Y = 1|H = h)$. The function $\mathcal{Z}(\cdot)$ depends on the genetic mode of the haplotype(s) associated with the disease. If the haplotype acts additively with respect to disease risk, then $\mathcal{Z}(H) = H^*$ where H^* is the haplotype-count vector H with the baseline haplotype element removed. If the haplotype acts dominantly, then $\mathcal{Z}(H) = I\{H^* \geq 1\}$, where the inequality is taken component wise, and $I\{A\} = 1$ if A is true. If the haplotype acts recessively, then $\mathcal{Z}(H) = I\{H^* = 2\}$. The vector β was taken to be the log of the vectors given in Table 2 for each simulation setting. The value of β_0 was set to maintain a disease prevalence between 3% and 5%. Once $P(Y = y|H = h)$ was calculated for each haplotype pair formed from the haplotype distributions given in Table 1, the vectors $P_{H|Y=y} = (P(H = h_1|Y = y) \cdots P(H = h_q|Y = y))$ were calculated for $Y = 0$ and $Y = 1$, where q is total number of haplotype pairs. The sample was generated by taking $n/2$ draws from the multinomial distribution parameterized by $P_{H|Y=0}$ to determine the haplotype pairs of the control individuals and by taking $n/2$ draws from the multinomial distribution parameterized by $P_{H|Y=1}$ to

determine the haplotype pairs of the case individuals. The haplotype pair of each individual was then dissolved into its unphased genotype.

COMPUTATIONAL DETAILS

For each simulation setting, 1,000 replicate datasets were generated, except for simulation under the null (i.e., $\theta = 1$), where 2,000 replicated datasets were generated. For each dataset, analysis began by calculating the unpenalized MLEs of the haplotype log-odds ratios. Prospective MLEs were obtained using `haplo.glm` in R [Lake et al, 2003] and retrospective MLEs were obtained using HAPSTAT in Linux [Lin et al, 2005]. The estimated covariance matrix of the MLEs was also obtained from each program. The final aPro and aRetro estimates were calculated by using the MLEs and their covariance matrix to compute the ALASSO solution via LSA. Based on these final estimates, estimation and model selection measures were calculated to compare the performance of the aPro and aRetro approaches. The estimation measures provided in this analysis are the bias and mean square error (MSE) of haplotype effect estimates. The model selection measures provided are the power to detect the associated haplotypes individually (referred to as *individual power*), the power to identify the true model (referred to as *true model power*), and the mean Type I error rate found by averaging the proportion of times a null haplotype was included in the model across all null haplotypes (referred to as *average type I error rate*).

RESULTS

We present the results of Null Simulation (i.e., no risk haplotypes) for both haplotype distributions in Table 3. For the simulations involving risk haplotypes, because the pattern of results was similar across both haplotype distributions, for brevity we focus the discussion on the

results for the first haplotype distribution. The results of Simulation I (single risk haplotype) for this setting are found in Tables 4 and 5. The results of Simulation II (two risk haplotypes) for this setting are found in Tables 6 and 7. The discussion generalizes to the second haplotype distribution, and specific results are shown in Tables 8 – 11. For each simulation, the results are broken down into two broad categories – correct analysis versus additive analysis. Correct analysis refers to specifying the correct genetic model when analyzing the data using haplo.glm or HAPSTAT, while additive analysis refers to analyzing non-additive data additively.

NULL SIMULATION

For each haplotype distribution, both aPro and aRetro have desirable and nearly identical performances under the null model (Table 3). The average type I error rate is low (ranging from 0.002 to 0.010) for both methods, and the true model power is high (ranging from 0.915 to 0.988). The effect estimations are also very similar, with the bias from both methods ranging from -0.001 to 0.007, and MSE ranging from 0.000 to 0.007. When comparing results between the two haplotype distributions, the true model power was higher and the average type I error rate was lower for Distribution 1 than Distribution 2. This result is not unexpected, as the dimension of Distribution 2 is larger and therefore more parameters need to be estimated. With the same amount of data to estimate more haplotype effects, the true model power decreases and the Type I error rates increases for Distribution 2.

SIMULATION I for Haplotype Distribution 1

Additive Genetic Mode When the risk haplotype acts additively, the individual power of each procedure is comparable. The individual power of aPro is within 5% of the individual power of

aRetro. If the effect size of the risk haplotype increases or if the frequency of the risk haplotype increases, the individual power of each procedure increases, but the relative power of the two methods stays the same. Similar results are observed when comparing the true model power and average Type I error rates of aPro and aRetro. Under this genetic model, the two procedures perform comparably with respect to all three model selection measures (Table 4).

When comparing the bias and MSE of aPro and aRetro, the relative performance of the two procedures hovers around 1, which indicates that aPro and aRetro perform similarly with respect to effect estimation (Table 4). For both procedures, the bias on the effect estimates is negative and the magnitudes are larger than what have been reported for an unpenalized likelihood analysis [e.g. Lin and Zeng, 2006]. These results are not unexpected when using a penalized likelihood approach. When a haplotype is not included in the model, its effect estimate is shrunk towards zero or set to exactly zero. Shrinkage can cause a large bias on the effect estimates. The impact of using a penalized method on the bias is greatest when the effect size is large and the power to detect the risk haplotype is low. As a result, a decrease in effect size or an increase in power does not necessarily guarantee a reduction in bias; the magnitude of the bias is a compromise between these two factors. This phenomenon is seen when examining the biases in Table 4: (1) the bias for rare risk haplotypes is larger than the corresponding common risk haplotypes with the same effect size because the power to detect rare risk haplotype is smaller; and (2) for a given risk haplotype, the bias on the effect estimate increases as the effect size increases until the power to detect the risk haplotype becomes large enough to overcome the shrinkage, and the bias on the effect estimate begins to decrease.

A similar pattern is observed when examining the MSE of the two procedures. Again, the MSE of aPro and aRetro is larger than what has been found in an unpenalized likelihood

analysis. MSE is an estimation measure that incorporates both the variance of an estimator and its bias. Because the effect estimates obtained from penalized methods are typically more efficient than those obtained from the corresponding unpenalized methods, it appears that the MSE of the effect estimates from aPro or aRetro could be dominated by their biases.

Dominant Genetic Mode Under a dominant genetic mode, aRetro performs slightly better than aPro when the data are analyzed under the correct genetic model (Table 4). The individual power of aRetro is between 1% and 11% higher than the individual power of aPro. Regardless of the frequency of the risk haplotype, the individual power gained by aRetro decreases as the effect size of the risk haplotype increases. Similar results are observed when comparing the true model power of the procedures, where the true model power of aRetro is between 1% and 13% higher than that of aPro. When comparing the bias and MSE of effect estimates from both procedures, the bias and MSE of aRetro is between 1% and 10% lower and between 1% and 17% lower, respectively, than that of aPro. Again, the gain in the performance of these measures decreases as the effect size of the associated haplotype increases.

When the data are incorrectly analyzed additively (Table 5), the average type I error rates are almost identical for both methods, and stay at a similar level as those of correct analysis. However, both methods suffer from a decrease in individual power (and hence the true model power), and an increase in bias and MSE on the effect estimates when compared to the performance of a correct analysis. It appears that the impact of using an incorrect genetic model is larger on aRetro than on aPro. For example, the aRetro power of identifying a common dominant risk haplotype with OR=1.7 is reduced from 0.794 to 0.699 under additive analysis, while the power reduction of aPro is from 0.737 to 0.709. The aRetro bias in this setting

increases from $|-0.145|$ to $|-0.242|$, while the bias of aPro increases from $|-0.161|$ to $|-0.225|$. As a result, aPro performs worse than aRetro under correct analysis (Table 4) but is comparable or slightly better than aRetro under the additive analysis (Table 5).

Recessive Genetic Mode Under a recessive genetic model, aRetro clearly outperforms aPro when the data are analyzed under the correct model (Table 4). When the risk haplotype is rare, the individual power of aRetro is at least 9 times higher than that of aPro, which essentially has no individual power under this scenario. When the risk haplotype is common, the individual power of aRetro is at least 1.5 times higher than that of aPro. Similar results are observed when comparing the true model power of the two procedures. The lack of power of aPro also manifests in substantial bias on the effect estimates. Bias from aRetro is at least 16% smaller than the bias of the effects estimates from aPro. MSE of the effect estimates from aRetro is at least 10% smaller than the MSE of the effects estimates from aPro.

When this data are incorrectly analyzed additively (Table 5), the performance of each method suffers from a decrease in power and an increase in bias and MSE on the effect estimates. The magnitude of the performance loss due to incorrect modeling is more severe than what was observed under a dominant model and is much more severe for aRetro than aPro. For example, the aRetro power of identifying a common recessive risk haplotype with $OR=2.5$ is reduced from 0.65 to 0.11 under additive analysis, while the power reduction of aPro is from 0.265 to 0.11. The aRetro bias in this setting increases from $|-0.329|$ to $|-0.877|$, while the bias of aPro increases from $|-0.549|$ to $|-0.876|$. Consequently, while aRetro exhibits absolute superiority over the aPro method (Table 4) under a correct analysis, it becomes comparable to or slightly better than aPro under the additive analysis (Table 5).

SIMULATION II for Haplotype Distribution 1

Simulation II examines the performance of aPro and aRetro when a two haplotypes are associated with the disease (Tables 6 and 7). Under each genetic mode, the pattern of results observed in Simulation I remain the same in Simulation II: (1) When the risk haplotypes act additively on disease susceptibility, the performances of aPro and aRetro are nearly identical for power, average type I error rate, bias and MSE. (2) The performance of aRetro is better than that of aPro under a dominant model with correct analysis. The gain brought by aRetro is the largest for two rare risk haplotypes, i.e., the most difficult scenario. However, when analyzing the data with an additive model, the power of both methods drops and the bias/MSE of both methods increase. The performance lose is more severe in aRetro, resulting in a comparable performance of aPro and aRetro. (3) Under a recessive model, aRetro has substantial power gain (at least 2 to 3 times higher) and smaller bias (e.g., can have 50% less bias) compared to aPro. However, when recessively acting haplotypes are analyzed using an additive genetic model, the performance of each procedure suffers, especially for aRetro. Both methods essentially lost their power to detect risk haplotypes and yielded sizable biases/MSE with the misspecification of the genetic mode.

SIMULATION I and II for Haplotype Distribution 2

When comparing results between the two haplotype distributions for Simulation I and II, individual power and true model power were typically higher for Distribution 1 while the average Type I error rate, bias and MSE were typically higher for Distribution 2. Like for the Null Simulation, these results are not unexpected because Distribution 2 has a larger dimension, which means more parameters needed to be estimated in the analysis. Increasing the number of

parameters and using the same amount of data for estimation can decrease power and increase bias. Although the magnitudes of model selection and estimation measures differ between the two haplotype distributions, when comparing the relative performance of aPro and aRetro, the pattern of results observed in the first haplotype distribution is similar to that in the second haplotype distribution for both Simulation I and II (Tables 8 – 11). The relative performance of aPro and aRetro depends on both the underlying genetic mode of the risk haplotypes and the genetic model adopted in the analysis. When the haplotypes associated with disease risk act additively, the two procedures perform comparably with respect to model selection and estimation measures. Under a dominant mode, aRetro performs slightly better than aPro and substantially better under a recessive mode. When these data are analyzed using an additive model, both procedures suffer from a loss in power and an increase in bias/MSE. The impact of imposing the incorrect genetic model is more severe for aRetro than for aPro, and the performance gain of aRetro is lost.

DISCUSSION

Like other haplotype-based methods developed to assess haplotype-phenotype association, the success of penalized regression methods depends on the underlying data likelihood. For unpenalized methods, prospective analyses are valid but less efficient than their retrospective counterparts for hypothesis testing, and can result in substantial bias when estimating the haplotype effects. Based on our simulation studies, the same can be said for penalized methods, which combine testing and estimation into one procedure. We found that the impact of using a prospective likelihood in the analysis depends on the underlying genetic mode of the associated genetic variant and the genetic model adopted in the analysis. When the genetic mode of the haplotypes is known and the correct inheritance model is imposed, using a

prospective analysis in place of the more appropriate retrospective analysis is detrimental when the associated haplotypes act dominantly or recessively with respect to disease risk. These results agree with the findings for non-penalized likelihood methods [Satten and Epstein, 2004]. Because the genetic mode of a genetic variant is usually unknown, researchers often analyze the data additively. When the dominant or recessive data are analyzed under an additive genetic model, the performance of the prospective and retrospective analyses become comparable: Both methods suffer from decreased power and increased bias for using an incorrect genetic model, and the retrospective analysis appears to be more sensitive to model misspecification and exhibits a larger degree of performance loss, thus making its performance gain over the prospective analysis negligible or slight.

While our simulations focused on penalized methods using the ALASSO penalty with the prospective likelihood of `haplo.glm` and the retrospective likelihood of `HAPSTAT`, we hope our findings can provide insight when coupling other penalized approaches with a prospective or a retrospective likelihood for case-control studies. If the main consideration is the relative performance of the retrospective vs. prospective penalized method, then our results suggest that the negative impact of developing haplotype-based penalized methods based on a prospective likelihood for case-control data is non-marginal only when the risk haplotypes act non-additively and the correct genetic model is adopted in the analysis. However, we think a more appropriate way to summarize our findings is to note that a careful haplotype-based penalized analysis of case-control data requires the use of a retrospective likelihood and the correct genetic mode. In practice, a major concern about using retrospective likelihoods is that they are difficult to implement. When using penalized likelihood methods, optimizing the retrospective likelihood can become even more intractable when the penalty term is incorporated. To overcome the

computational burden, the least squares approximation may provide a promising alternative for implementing penalized retrospective methods. By using the least square approximation, the need to directly optimize the penalized retrospective likelihood is bypassed. Instead, the unpenalized likelihood is optimized once for a starting value in the approximation; hence implementing prospective and retrospective penalized methods have similar computational costs. The spared computational efforts can be put into exploring and identifying the correct genetic mode for potential risk haplotypes.

Penalized likelihood methods can have higher power than unpenalized methods in detecting important haplotypic factors [Guo and Lin, 2009]. Our simulations also reveal that the methods considered here can have better power to identify risk haplotypes individually and have better power to identify the true model than the unpenalized version (results not shown). This means that the penalized likelihood approach often identifies the truly associated haplotypes and only the truly associated haplotypes. However, the power enjoyed by penalized likelihood methods comes at the expense of obtaining effect estimates with higher bias than their unpenalized likelihood counterparts. As observed in our simulations, the bias on the effect estimates obtained by the penalized method can remain sizeable even when the power to detect the effects is reasonably high. In addition to power, the bias can be affected by the model selection criterion used to select the tuning parameter. In our analysis, we used BIC to choose the final model because it can achieve consistent model selection [Shao, 1997; Yang, 2005] and our goal was to identify the true model structure. To achieve selection consistency, BIC penalizes degrees of freedom more heavily, which can place a larger amount of shrinkage on the effect estimates and increase their bias. Alternatively, a cross-validation criterion or AIC could be used to select the final model. These selection criteria target prediction error rather than finding the

true model structure [Shao, 1997; Yang, 2005] and typically impose a smaller penalty of degrees of freedom than BIC. As a result, models selected using these criteria incur less shrinkage on the effect estimates which can decrease their bias, but also can increase the chance of including non-important predictors in the final model.

ACKNOWLEDGEMENTS

MLK was supported by NIH T32GM081057. HDB was supported by NSF DMS-0705968 and NIH R01 MH084022-01. JYT was supported by NIH R01 MH084022-01. The authors thank Dr. Danyu Lin, Tammy Bailey, and Chris Smith for their generous help with HAPSTAT and perl.

REFERENCES

- Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7: 781 – 791.
- Clark AG. 2004. The role of haplotypes in candidate-gene studies. *Genet Epidemiol* 27: 321 – 333.
- Cordell HJ. 2006. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genet Epidemiol* 30: 259 – 275.
- de Bakker PW, Yelensky R, Pe'er I, Gabriel S, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* 37: 1217 – 1223.
- Guo W, Lin S. 2009. Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epidemiol* 33: 308 – 316.
- Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 55: 56 – 65.

- Li Y, Sung WK, Liu JJ. 2007. Association mapping via regularized regression analysis of single nucleotide polymorphism haplotypes in variable-sized sliding windows. *Am J Hum Genet* 80: 705 – 715.
- Lin DY, Huang BE. 2008. The use of inferred haplotypes in downstream analysis. *Am J Hum Genet* 80: 577 – 579.
- Lin DY, Zeng D. 2006. Likelihood-based inference on haplotype effects in genetic association studies. *JASA* 101: 89 – 104.
- Lin DY, Zeng D, Milikan R. 2005. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genet Epidemiol* 29: 299 – 312.
- Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66: 403 – 412.
- Satten GA, Epstein MP. 2004. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol* 27: 192 – 201.
- Schaid DJ. 2004. Evaluating associations of haplotypes with traits. *Genet Epidemiol* 27: 348 – 364.
- Shao J. 1997. An asymptotic theory for linear model selection. *Stat Sinica* 7: 221 – 264.
- Souverein OW, Zwinderman AH, Tanck MW. 2006. Estimating haplotype effects on dichotomous outcome for unphased genotype data using a weighted penalize log-likelihood approach. *Hum Hered* 61: 104 – 110.
- Souverein OW, Zwinderman AH, Jukema JW, Tanck MW. 2008. Estimating effects of rare haplotypes on failure times using a penalized Cox proportional hazards regression model. *BMC Genet* 9: 9.
- Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003. Modeling and E-M estimation of haplotype specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered* 55:179-190.
- Tanck MW, Klerkx AH, Jukema JW, De Knijff P, Kastelein JJ, Zwinderman AH. 2003. Estimation of multilocus haplotype effects using weighted penalized log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. *Ann Hum Genet* 67: 175 – 184.

- Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. *J R Stat Soc* 58: 267 – 288.
- Tzeng JY, Bondell HD. 2010. A comprehensive approach to haplotype-specific analysis by penalized likelihood. *Euro J Hum Genet* 18: 95 – 103.
- Tzeng JY, Lu W, Farmen MW, Liu Y, Sullivan PF. 2010. Haplotype-based pharmacogenetic analysis for longitudinal quantitative traits in the presence of dropout. *J of Biopharm Stats*: In Press.
- Wang H, Leng C. 2007. Unified LASSO estimation by least squares approximation. *J Am Stat Assoc* 102: 1039 – 1048.
- Yang Y. 2005. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92: 937 – 950.
- Zaitlen M, Kang H, Eskin E, Halperin E. 2007. Leveraging the HapMap correlation structure in association studies. *Am J Hum Genet* 80: 683 – 691.
- Zou H. 2006. The adaptive LASSO and its oracle properties. *J Am Stat Assoc* 101: 1418 – 1429.

Table 1: Haplotype distributions used in simulation studies

Hap ID	Distribution 1		Distribution 2	
	Haplotype	Frequency	Haplotype	Frequency
1	00000	0.406	00010	0.131
2	00001	0.213	00001	0.105
3	01111	0.141	10010	0.103
4	10000	0.132	10101	0.100
5	10001	0.055	00100	0.088
6	01000	0.021	10100	0.088
7	01100	0.018	00101	0.086
8	01001	0.014	01101	0.084
9			10001	0.081
10			10000	0.079
11			00000	0.055

Table 2: Settings for Simulation I and Simulation II (odds ratios)

Hap ID	Distribution 1						Distribution 2					
	Freq	<u>Sim I</u>		<u>Sim II</u>			Freq	<u>Sim I</u>		<u>Sim II</u>		
		R	C	R/R	R/C	C/C		R	C	R/R	R/C	C/C
1	0.406	*	*	*	*	*	0.131	*	*	*	*	*
2	0.213	1	1	1	1	1	0.105	1	θ	1	θ	θ
3	0.141	1	1	1	1	θ	0.103	1	1	1	1	θ
4	0.132	1	θ	1	θ	θ	0.100	1	1	1	1	1
5	0.055	θ	1	θ	θ	1	0.088	1	1	1	1	1
6	0.021	1	1	θ	1	1	0.088	1	1	1	1	1
7	0.018	1	1	1	1	1	0.086	1	1	1	1	1
8	0.014	1	1	1	1	1	0.084	1	1	1	1	1
9							0.081	1	1	1	1	1
10							0.079	1	1	θ	1	1
11							0.055	θ	1	θ	θ	1

Table 3: Results of Null Simulation

		<i>Model Selection Results</i>						<i>Parameter Estimation Results</i>						
		True Model Power			Average Type I Error Rate			Bias			MSE			
		Model*	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P
Haplotype Distribution 1	<i>Correct Analysis</i>	Additive	0.976	0.974	1.00	0.004	0.004	1.00	0.000	-0.001	–	0.002	0.002	1.00
		Dominant	0.978	0.971	0.99	0.004	0.005	1.25	0.000	0.000	–	0.002	0.003	1.50
		Recessive	0.988	0.978	0.99	0.002	0.004	2.00	-0.001	0.003	-3.00	0.001	0.003	3.00
Haplotype Distribution 2	<i>Correct Analysis</i>	Additive	0.915	0.950	1.04	0.010	0.006	0.60	0.000	0.000	–	0.002	0.002	1.00
		Dominant	0.925	0.968	1.05	0.008	0.004	0.50	0.001	0.001	1.00	0.001	0.001	1.00
		Recessive	0.938	0.955	1.02	0.007	0.005	0.71	0.000	0.007	–	0.000	0.007	–

* Model refers to genetic model adopted in the analysis.

Table 4: Results of Simulation I for Haplotype Distribution 1 – Correct Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power				True Model Power			Average Type I Error Rate			Bias			MSE		
Mode*	Freq	OR	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P
Additive	<i>Rare</i>	1.3	0.057	0.055	0.96	0.052	0.048	0.92	0.003	0.004	1.33	-0.230	-0.231	1.00	0.071	0.071	1.00
		1.5	0.180	0.195	1.08	0.175	0.190	1.09	0.002	0.003	1.50	-0.304	-0.298	0.98	0.142	0.140	0.99
		1.7	0.377	0.386	1.02	0.343	0.351	1.02	0.008	0.008	1.00	-0.305	-0.304	1.00	0.186	0.183	0.98
		2.0	0.756	0.761	1.01	0.700	0.702	1.00	0.010	0.011	1.10	-0.201	-0.200	1.00	0.144	0.142	0.99
	<i>Common</i>	1.3	0.199	0.208	1.05	0.183	0.193	1.05	0.005	0.004	0.80	-0.190	-0.188	0.99	0.059	0.058	0.98
		1.5	0.535	0.540	1.01	0.470	0.470	1.00	0.013	0.013	1.00	-0.189	-0.187	0.99	0.082	0.081	0.99
		1.7	0.880	0.888	1.01	0.828	0.830	1.00	0.011	0.012	1.09	-0.103	-0.101	0.98	0.050	0.048	0.96
		2.0	0.997	0.997	1.00	0.947	0.946	1.00	0.009	0.009	1.00	-0.071	-0.072	1.01	0.027	0.027	1.00
Dominant	<i>Rare</i>	1.3	0.054	0.059	1.09	0.050	0.053	1.06	0.003	0.005	1.67	-0.230	-0.228	0.99	0.072	0.071	0.99
		1.5	0.130	0.160	1.23	0.120	0.150	1.25	0.004	0.003	0.60	-0.324	-0.311	0.96	0.151	0.146	0.97
		1.7	0.341	0.368	1.08	0.315	0.345	1.10	0.006	0.005	0.83	-0.317	-0.307	0.97	0.196	0.188	0.96
		2.0	0.668	0.697	1.04	0.616	0.636	1.03	0.011	0.012	1.09	-0.251	-0.239	0.95	0.184	0.170	0.92
	<i>Common</i>	1.3	0.136	0.151	1.11	0.128	0.144	1.13	0.004	0.004	1.00	-0.207	-0.203	0.98	0.063	0.062	0.98
		1.5	0.410	0.445	1.09	0.370	0.405	1.10	0.010	0.009	0.92	-0.231	-0.224	0.97	0.101	0.096	0.95
		1.7	0.737	0.794	1.08	0.681	0.734	1.08	0.011	0.011	1.00	-0.161	-0.145	0.90	0.088	0.073	0.83
		2.0	0.965	0.975	1.01	0.907	0.917	1.01	0.010	0.011	1.10	-0.090	-0.091	1.01	0.046	0.041	0.89
Recessive	<i>Rare</i>	2.0	0.010	0.090	9.00	0.000	0.085	–	0.002	0.006	3.00	-0.693	-0.583	0.84	0.480	0.474	0.99
		2.5	0.010	0.130	13.00	0.000	0.120	–	0.002	0.002	1.00	-0.916	-0.753	0.82	0.840	0.753	0.90
		3.0	0.010	0.230	23.00	0.000	0.205	–	0.002	0.002	1.00	-1.099	-0.764	0.70	1.207	0.987	0.82
		3.5	0.005	0.280	56.00	0.000	0.275	–	0.001	0.001	1.00	-1.253	-0.865	0.69	1.569	1.163	0.74
	<i>Common</i>	2.0	0.075	0.305	4.07	0.075	0.295	3.93	0.000	0.002	–	-0.603	-0.441	0.73	0.467	0.347	0.74
		2.5	0.265	0.650	2.45	0.255	0.630	2.47	0.002	0.005	2.50	-0.549	-0.329	0.60	0.711	0.321	0.45
		3.0	0.520	0.890	1.71	0.505	0.850	1.68	0.002	0.008	4.00	-0.440	-0.226	0.51	0.631	0.188	0.30
		3.5	0.645	0.940	1.46	0.615	0.920	1.50	0.005	0.003	0.60	-0.367	-0.242	0.66	0.671	0.176	0.26

* Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype(s); Labeling holds for Tables 4 – 11.

Table 5: Results of Simulation I for Haplotype Distribution 1 – Additive Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power			True Model Power			Average Type I Error Rate			Bias			MSE			
Mode	Freq	OR	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P
Dominant	<i>Rare</i>	1.3	0.051	0.044	0.86	0.045	0.039	0.87	0.004	0.003	0.75	-0.234	-0.237	1.01	0.070	0.070	1.00
		1.5	0.135	0.140	1.04	0.120	0.125	1.04	0.004	0.004	1.00	-0.326	-0.324	0.994	0.149	0.148	0.99
		1.7	0.335	0.323	0.96	0.303	0.291	0.96	0.007	0.007	1.00	-0.331	-0.342	1.03	0.195	0.197	1.01
		2.0	0.644	0.634	0.98	0.582	0.564	0.97	0.012	0.014	1.17	-0.286	-0.304	1.06	0.194	0.198	1.02
	<i>Common</i>	1.3	0.121	0.117	0.97	0.113	0.11	0.97	0.004	0.004	1.00	-0.219	-0.221	1.01	0.062	0.062	1.00
		1.5	0.350	0.340	0.97	0.315	0.310	0.98	0.010	0.008	0.83	-0.273	-0.281	1.029	0.110	0.111	1.01
		1.7	0.709	0.699	0.99	0.655	0.643	0.98	0.011	0.011	1.00	-0.225	-0.242	1.08	0.099	0.103	1.04
		2.0	0.951	0.949	1.00	0.875	0.87	0.99	0.014	0.015	1.07	-0.186	-0.216	1.16	0.068	0.077	1.13
Recessive	<i>Rare</i>	2.0	0.015	0.015	1.00	0.005	0.050	10.00	0.004	0.006	1.50	-0.688	-0.689	1.00	0.481	0.481	1.00
		2.5	0.000	0.005	–	0.000	0.030	–	0.004	0.004	1.00	-0.916	-0.914	1.00	0.840	0.836	1.00
		3.0	0.015	0.020	1.33	0.005	0.040	8.00	0.004	0.005	1.25	-1.089	-1.088	1.00	1.192	1.189	1.00
		3.5	0.000	0.010	–	0.000	0.015	–	0.000	0.001	–	-1.253	-1.248	1.00	1.569	1.561	0.99
	<i>Common</i>	2.0	0.025	0.050	2.00	0.020	0.030	1.50	0.001	0.003	3.00	-0.685	-0.684	1.00	0.472	0.470	1.00
		2.5	0.110	0.110	1.00	0.105	0.110	1.05	0.002	0.004	2.00	-0.876	-0.877	1.00	0.781	0.782	1.00
		3.0	0.145	0.155	1.07	0.140	0.155	1.11	0.002	0.005	2.50	-1.051	-1.044	0.99	1.120	1.107	0.99
		3.5	0.260	0.285	1.10	0.260	0.275	1.06	0.002	0.003	1.50	-1.156	-1.146	0.99	1.365	1.343	0.98

Table 6: Results of Simulation II for Haplotype Distribution 1 – Correct Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power			True Model Power			Average Type I Error Rate			Bias			MSE			
Mode	Freq	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	
Additive	R/R	R1	0.105	0.106	1.01	0.078	0.078	1.00	0.006	0.006	1.00	-0.443	-0.444	1.00	0.267	0.266	1.00
		R2	0.347	0.358	1.03							-0.327	-0.324	0.99	0.193	0.19	0.98
	R/C	R1	0.497	0.5	1.01	0.425	0.419	0.99	0.018	0.02	1.11	-0.25	-0.249	1.00	0.156	0.155	0.99
		C1	0.907	0.912	1.01							-0.109	-0.109	1.00	0.048	0.046	0.96
	C/C	C1	0.878	0.874	1.00	0.77	0.772	1.00	0.019	0.018	0.95	-0.113	-0.115	1.02	0.053	0.054	1.02
		C2	0.918	0.921	1.00							-0.101	-0.103	1.02	0.042	0.042	1.00
Dominant	R/R	R1	0.104	0.124	1.19	0.066	0.083	1.26	0.007	0.008	1.14	-0.441	-0.433	0.98	0.27	0.261	0.97
		R2	0.268	0.31	1.16							-0.371	-0.352	0.95	0.214	0.202	0.94
	R/C	R1	0.464	0.48	1.03	0.37	0.381	1.03	0.015	0.017	1.13	-0.253	-0.252	1.00	0.167	0.162	0.97
		C1	0.795	0.816	1.03							-0.142	-0.133	0.94	0.077	0.071	0.92
	C/C	C1	0.736	0.801	1.09	0.615	0.66	1.07	0.017	0.023	1.35	-0.17	-0.149	0.88	0.092	0.074	0.80
		C2	0.787	0.856	1.09							-0.152	-0.13	0.86	0.076	0.059	0.78
Recessive	R/R	R1	0.000	0.000	–	0.000	0.000	–	0.001	0.004	4.00	-0.916	-0.916	1.00	0.84	0.84	1.00
		R2	0.000	0.135	–							-0.916	-0.731	0.80	0.84	0.766	0.91
	R/C	R1	0.010	0.170	17.00	0.005	0.135	27.00	0.002	0.006	3.00	-0.916	-0.684	0.75	0.84	0.749	0.89
		C1	0.225	0.615	2.73							-0.618	-0.367	0.59	0.72	0.354	0.49
	C/C	C1	0.205	0.655	3.20	0.165	0.510	3.09	0.000	0.010	–	-0.662	-0.319	0.48	0.717	0.322	0.45
		C2	0.305	0.750	2.46							-0.547	-0.292	0.53	0.664	0.244	0.37

Table 7: Results of Simulation II for Haplotype Distribution 1 – Additive Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power			True Model Power			Average Type I Error Rate			Bias			MSE			
Mode	Freq	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	
Dominant	R/R	R1	0.100	0.099	0.99	0.065	0.062	0.95	0.006	0.006	1.00	-0.448	-0.452	1.01	0.268	0.265	0.99
		R2	0.247	0.241	0.98							-0.391	-0.397	1.02	0.218	0.219	1.00
	R/C	R1	0.470	0.478	1.02	0.358	0.369	1.03	0.016	0.017	1.06	-0.261	-0.265	1.02	0.163	0.160	0.98
		C1	0.740	0.733	0.99							-0.216	-0.233	1.08	0.094	0.098	1.04
	C/C	C1	0.675	0.667	0.99	0.551	0.545	0.99	0.017	0.019	1.12	-0.243	-0.255	1.05	0.110	0.113	1.03
		C2	0.731	0.730	1.00							-0.225	-0.237	1.05	0.096	0.098	1.02
Recessive	R/R	R1	0.000	0.000	–	0.000	0.000	–	0.003	0.005	1.67	-0.916	-0.916	1.00	0.840	0.840	1.00
		R2	0.000	0.000	–							-0.916	-0.916	1.00	0.840	0.840	1.00
	R/C	R1	0.020	0.015	0.75	0.010	0.010	1.00	0.005	0.006	1.20	-0.916	-0.916	1.00	0.843	0.844	1.00
		C1	0.080	0.080	1.00							-0.885	-0.884	1.00	0.795	0.794	1.00
	C/C	C1	0.060	0.065	1.08	0.025	0.025	1.00	0.003	0.005	1.67	-0.895	-0.894	1.00	0.809	0.807	1.00
		C2	0.065	0.070	1.08							-0.895	-0.894	1.00	0.808	0.806	1.00

Table 8: Results of Simulation I for Haplotype Distribution 2 – Correct Analysis

		<i>Model Selection Results</i>										<i>Parameter Estimation Results</i>					
		Individual Power				True Model Power			Average Type I Error Rate			Bias			MSE		
Mode	Freq	OR	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P
Additive	<i>Rare</i>	1.3	0.035	0.040	1.14	0.030	0.030	1.00	0.007	0.004	0.57	-0.245	-0.244	1.00	0.068	0.068	1.00
		1.5	0.115	0.135	1.17	0.110	0.130	1.18	0.010	0.005	0.50	-0.338	-0.329	0.97	0.151	0.148	0.98
		1.7	0.325	0.315	0.97	0.270	0.270	1.00	0.009	0.008	0.89	-0.349	-0.351	1.01	0.198	0.201	1.02
		2.0	0.710	0.710	1.00	0.600	0.640	1.07	0.016	0.009	0.56	-0.250	-0.250	1.00	0.169	0.168	0.99
	<i>Common</i>	1.3	0.285	0.290	1.02	0.265	0.275	1.04	0.009	0.008	0.89	-0.229	-0.221	0.97	0.067	0.067	1.00
		1.5	0.615	0.685	1.11	0.570	0.560	0.98	0.006	0.006	1.00	-0.311	-0.286	0.92	0.130	0.121	0.93
		1.7	0.875	0.880	1.01	0.820	0.835	1.02	0.007	0.009	1.29	-0.443	-0.382	0.86	0.233	0.199	0.85
		2.0	0.995	0.990	0.99	0.920	0.940	1.02	0.010	0.018	1.80	-0.646	-0.583	0.90	0.450	0.404	0.90
Dominant	<i>Rare</i>	1.3	0.045	0.070	1.56	0.040	0.060	1.50	0.009	0.004	0.44	-0.234	-0.219	0.94	0.073	0.076	1.04
		1.5	0.105	0.120	1.14	0.095	0.100	1.05	0.009	0.005	0.56	-0.343	-0.336	0.98	0.152	0.151	0.99
		1.7	0.230	0.310	1.35	0.175	0.235	1.34	0.014	0.012	0.86	-0.388	-0.346	0.89	0.224	0.203	0.91
		2.0	0.490	0.575	1.17	0.400	0.515	1.29	0.014	0.009	0.64	-0.361	-0.322	0.89	0.260	0.225	0.87
	<i>Common</i>	1.3	0.185	0.200	1.08	0.180	0.206	1.14	0.003	0.004	1.33	-0.240	-0.234	0.98	0.067	0.068	1.01
		1.5	0.455	0.480	1.05	0.425	0.430	1.01	0.008	0.012	1.50	-0.296	-0.273	0.92	0.132	0.123	0.93
		1.7	0.730	0.765	1.05	0.655	0.670	1.02	0.008	0.011	1.38	-0.376	-0.339	0.90	0.205	0.185	0.90
		2.0	0.980	0.990	1.01	0.885	0.880	0.99	0.012	0.014	1.17	-0.509	-0.441	0.87	0.349	0.304	0.87
Recessive	<i>Rare</i>	2.0	0.000	0.070	–	0.000	0.070	–	0.006	0.004	0.67	-0.693	-0.610	0.88	0.480	0.467	0.97
		2.5	0.000	0.190	–	0.000	0.155	–	0.006	0.008	1.33	-0.916	-0.660	0.72	0.840	0.728	0.87
		3.0	0.000	0.190	–	0.000	0.175	–	0.008	0.003	0.38	-1.099	-0.859	0.78	1.207	0.995	0.82
		3.5	0.000	0.315	–	0.000	0.285	–	0.005	0.006	1.20	-1.253	-0.841	0.67	1.569	1.103	0.70
	<i>Common</i>	2.0	0.065	0.180	2.77	0.065	0.165	2.54	0.003	0.007	2.33	-0.693	-0.539	0.78	0.481	0.404	0.84
		2.5	0.120	0.390	3.25	0.120	0.340	2.83	0.001	0.008	8.00	-0.856	-0.555	0.65	0.824	0.531	0.64
		3.0	0.300	0.535	1.78	0.300	0.490	1.63	0.002	0.007	3.50	-0.932	-0.548	0.59	1.087	0.590	0.54
		3.5	0.415	0.595	1.43	0.415	0.530	1.28	0.000	0.011	–	-0.974	-0.590	0.61	1.284	0.690	0.54

Table 9: Results of Simulation I for Haplotype Distribution 2 – Additive Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power			True Model Power			Average Type I Error Rate			Bias			MSE			
Mode	Freq	OR	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P
Dominant	<i>Rare</i>	1.3	0.050	0.045	0.90	0.040	0.040	1.00	0.011	0.005	0.45	-0.232	-0.235	1.01	0.073	0.072	0.99
		1.5	0.100	0.100	1.00	0.090	0.085	0.94	0.008	0.005	0.63	-0.351	-0.351	1.00	0.151	0.152	1.01
		1.7	0.205	0.210	1.02	0.165	0.170	1.03	0.014	0.008	0.57	-0.406	-0.408	1.00	0.229	0.227	0.99
		2.0	0.495	0.450	0.91	0.420	0.390	0.93	0.011	0.008	0.73	-0.390	-0.414	1.06	0.263	0.281	1.07
	<i>Common</i>	1.3	0.180	0.195	1.08	0.170	0.175	1.03	0.004	0.003	0.75	-0.243	-0.245	1.01	0.067	0.067	1.00
		1.5	0.435	0.430	0.99	0.400	0.410	1.03	0.008	0.008	1.00	-0.314	-0.308	0.98	0.133	0.129	0.97
		1.7	0.725	0.730	1.01	0.665	0.660	0.99	0.008	0.011	1.38	-0.397	-0.385	0.97	0.208	0.200	0.96
		2.0	0.985	0.980	0.99	0.915	0.930	1.02	0.010	0.011	1.10	-0.539	-0.478	0.89	0.358	0.304	0.85
Recessive	<i>Rare</i>	2.0	0.000	0.000	–	0.000	0.000	–	0.008	0.003	0.38	-0.693	-0.693	1.00	0.480	0.480	1.00
		2.5	0.005	0.005	1.00	0.005	0.005	1.00	0.012	0.006	0.50	-0.912	-0.912	1.00	0.835	0.835	1.00
		3.0	0.005	0.010	2.00	0.005	0.010	2.00	0.009	0.001	0.11	-1.096	-1.094	1.00	1.203	1.199	1.00
		3.5	0.015	0.015	1.00	0.010	0.010	1.00	0.008	0.004	0.50	-1.244	-1.242	1.00	1.552	1.551	1.00
	<i>Common</i>	2.0	0.070	0.050	0.71	0.065	0.040	0.62	0.004	0.002	0.50	-0.691	-0.691	1.00	0.478	0.478	1.00
		2.5	0.110	0.110	1.00	0.100	0.105	1.05	0.005	0.006	1.20	-0.903	-0.895	0.99	0.821	0.809	0.99
		3.0	0.210	0.195	0.93	0.210	0.190	0.90	0.004	0.002	0.50	-1.087	-1.078	0.99	1.186	1.171	0.99
		3.5	0.270	0.285	1.06	0.255	0.275	1.08	0.002	0.003	1.50	-1.230	-1.216	0.99	1.523	1.495	0.98

Table 10: Results of Simulation II for Haplotype Distribution 2 – Correct Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power			True Model Power			Average Type I Error Rate			Bias			MSE			
Mode	Freq	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	
Additive	R/R	R1	0.365	0.395	1.08	0.230	0.260	1.13	0.024	0.018	0.75	-0.318	-0.301	0.95	0.190	0.184	0.97
		R2	0.605	0.620	1.02							-0.227	-0.219	0.96	0.125	0.122	0.98
	R/C	R1	0.100	0.190	1.90	0.090	0.150	1.67	0.015	0.015	1.00	-0.470	-0.422	0.90	0.257	0.234	0.91
		C1	0.875	0.880	1.01							-0.417	-0.333	0.80	0.219	0.176	0.80
	C/C	C1	0.535	0.520	0.97	0.110	0.150	1.36	0.012	0.016	1.33	-0.467	-0.425	0.91	0.251	0.228	0.91
		C2	0.505	0.500	0.99							-0.452	-0.416	0.92	0.235	0.217	0.92
Dominant	R/R	R1	0.365	0.455	1.25	0.170	0.230	1.35	0.021	0.016	0.76	-0.319	-0.279	0.87	0.188	0.166	0.88
		R2	0.515	0.570	1.11							-0.248	-0.229	0.92	0.151	0.135	0.89
	R/C	R1	0.180	0.340	1.89	0.110	0.125	1.14	0.021	0.026	1.24	-0.414	-0.337	0.81	0.240	0.197	0.82
		C1	0.760	0.770	1.01							-0.391	-0.349	0.89	0.210	0.184	0.88
	C/C	C1	0.500	0.515	1.03	0.140	0.155	1.11	0.009	0.012	1.33	-0.456	-0.411	0.90	0.243	0.212	0.87
		C2	0.420	0.435	1.04							-0.434	-0.379	0.87	0.229	0.194	0.85
Recessive	R/R	R1	0.000	0.130	–	0.000	0.060	–	0.004	0.006	1.50	-0.916	-0.749	0.82	0.840	0.763	0.91
		R2	0.000	0.260	–							-0.916	-0.642	0.70	0.840	0.642	0.76
	R/C	R1	0.000	0.145	–	0.000	0.075	–	0.003	0.006	2.00	-0.916	-0.736	0.80	0.840	0.746	0.89
		C1	0.120	0.455	3.79							-0.846	-0.517	0.61	0.827	0.487	0.59
	C/C	C1	0.155	0.315	2.03	0.015	0.150	10.00	0.001	0.011	11.00	-0.885	-0.632	0.71	0.854	0.593	0.69
		C2	0.120	0.355	2.96							-0.874	-0.604	0.69	0.824	0.560	0.68

Table 11: Results of Simulation II for Haplotype Distribution 2 – Additive Analysis

		<i>Model Selection Results</i>									<i>Parameter Estimation Results</i>						
		Individual Power			True Model Power			Average Type I Error Rate			Bias			MSE			
Mode	Freq	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	Pro	Retro	R/P	
Dominant	R/R	R1	0.350	0.335	0.96	0.160	0.205	1.28	0.020	0.014	0.70	-0.337	-0.348	1.03	0.192	0.195	1.02
		R2	0.510	0.530	1.04							-0.286	-0.289	1.01	0.153	0.148	0.97
	R/C	R1	0.135	0.225	1.67	0.075	0.140	1.87	0.019	0.024	1.26	-0.443	-0.402	0.91	0.249	0.225	0.90
		C1	0.745	0.750	1.01							-0.413	-0.381	0.92	0.213	0.192	0.90
	C/C	C1	0.490	0.510	1.04	0.090	0.090	1.00	0.011	0.016	1.45	-0.469	-0.458	0.98	0.246	0.238	0.97
		C2	0.375	0.380	1.01							-0.458	-0.453	0.99	0.241	0.236	0.98
Recessive	R/R	R1	0.010	0.010	1.00	0.000	0.000	–	0.006	0.003	0.50	-0.911	-0.911	1.00	0.833	0.833	1.00
		R2	0.010	0.030	3.00							-0.913	-0.905	0.99	0.834	0.823	0.99
	R/C	R1	0.005	0.005	1.00	0.000	0.000	–	0.006	0.003	0.50	-0.913	-0.913	1.00	0.836	0.836	1.00
		C1	0.115	0.119	1.03							-0.902	-0.895	0.99	0.819	0.809	0.99
	C/C	C1	0.155	0.145	0.94	0.050	0.080	2.00	0.006	0.006	1.00	-0.910	-0.907	1.00	0.830	0.826	1.00
		C2	0.115	0.135	1.17							-0.906	-0.901	0.99	0.824	0.818	0.99