Published in final edited form as:

J Biopharm Stat. 2008 ; 18(6): 1150–1177. doi:10.1080/10543400802369038.

A Mixture Model Approach in Gene-Gene and Gene-Environmental Interactions for Binary Phenotypes

Lang Li, Ph.D., Menggang Yu, Ph.D., Robarge D. Jason, M.S., Changyu Shen, PhD, and Faouzi Azzouz, MS

Division of Biostatistics, Department of Medicine, Indiana University

Howard L. McLeod, PhD.

Department of Medicine, Washington University School of Medicine; St. Louis, Missouri

Silvana Borges-Gonzales, MD, Anne Nguyen, BS, Todd Skaar, PhD, Zeruesenay Desta, PhD, Christopher J. Sweeney, MD, and David A Flockhart, MD, PhD

Division of Clinical Pharmacology, Department of Medicine, Indiana University

Summary

In translational research, a genetic association study of a binary outcome has a two-fold aim: test whether genetic/environmental variables or their combinations are associated with a clinical phenotype; and determine how those combinations are grouped to predict the phenotype (i.e. which combinations have a similarly distributed phenotype, and which ones have differently distributed phenotypes). The second part of this aim has high clinical appeal, because it can directly facilitate clinical decisions. Although traditional logistic regression can detect gene-gene or gene-environmental interaction effects on binary phenotypes, they cannot decisively determine how genotype combinations are grouped to predict the phenotype. Our proposed mixture model approach is valuable in this context. It concurrently detects main and interaction effects of genetic and environmental variables through a likelihood ratio test (LRT), and conducts phenotype cluster analysis based on genetic and environmental variable combinations. The theoretical distribution of the proposed mixture model's likelihood ratio test is robust not only to small sample size, but also to unequal sample size in various genotype and environmental subgroups. Hypothesis testing through a likelihood ratio test results in a fast algorithm for p-value calculations. Extensive simulation studies demonstrate that mixture model, overall test in logistic regression, and Monte Carlo based logic regression constantly possess the best power to detect multi-way gene/environmental combinations. The mixture model approach has the highest recovery probability to recover the true partition in the simulation studies. Its applications are exemplified in interim data analyses for two cancer studies.

Keywords

gene-gene interaction; mixture model; pharmacogenetics

1. Introduction

1.1 Background and Significance

As complex disease and drug response are recognized to involve both genetic and environmental factors, genetic studies focus not only on the effects of genetic variation in

candidate genes, but also on the potential interactions between genetic and environmental variables. A pharmacogenetic study has two related goals: test for associations between candidate genes' variations and the results of drug therapy; and utilize combinations of genetic polymorphisms present in drug specific metabolizing enzymes, transporters, and pharmacodynamic receptors to predict both intermediate phenotypes (e.g. pharmacokinetic parameters: drug clearance, parent drug to metabolite ratio, etc.), and clinical outcome phenotypes (i.e. disease free survival, and side effects). Interactions influencing a phenotype are not limited to genetic factors. Gene-environment interactions occur when an environmental factor mediates the expressions of genetic factors. If the effect of a measured genotype on a phenotype varies with environment, the genetic effect may be detectable only when accounting for the environmental factor(s) (e.g. race, gender, menopausal status, and concomitant medications).

Our research demonstrates the value of a pharmacogenetic approach to understand both genetic and environmental effects on the treatment outcomes in various diseases. In a study of QT prolongation syndrome due to haloperidol, prolongation of the electrocardiographic QT interval was significant only in males; and only in male group, subjects with CYP2D6 poor metabolizer genotypes tended to have larger QT prolongation than subject with CYP2D6 extensive metabolizer genotype [1]. In a study of hypertension, an association between CYP3A5 genotype and the baseline blood pressure was significant in African American, where subjects with CYP3A5 *1/*1 genotype had higher BP than subjects with *1/*3 or *3/*3. However, this association was not detected in Caucasian samples [2]. In our breast cancer studies, genetic polymorphisms in the enzyme responsible for metabolism of tamoxifen to its active metabolite, the cytochrome P450 isoform CYP2D6, as well as environmental variables (i.e. concomitant medications), and their interactions have been shown to predict pharmacokinetic [3-5]. In the same breast cancer study, the genetic effects of estrogen receptor 1 (ESR-1) and estrogen receptor 2 (ESR-2) on the lipid change after tamoxifen treatment were significant only among breast cancer patients who did not take lipid lower medications. Other gene-gene and gene-environment interaction examples are well documented [6,7].

1.2 Challenges in Gene/Gene and Gene/Environmental Interactions

Consider a binary phenotype related to two, biallelic contributing genetic loci. Let the binary phenotype be interpreted as risk in our context. Denote the alleles from locus A be denoted "A" and "a", and denote those from locus B denoted "B" and "b". At locus A, there are three genotypes: (AA, Aa, aa); and at locus B, three genotypes are (BB, Bb, bb). Figure 1 displays their 9 combination cells. Figure 1 (a) illustrates the risk distribution under the null hypothesis, in which all cells share the same disease risk, p_0 . Figure 1 (b) presents a main effect from loci B. Genotype bb has risk p_1 , while genotypes, (BB, Bb), have p_2 . Figure 1 (c) demonstrates an interaction between loci A and B. The diagonal cells have risk of p_1 , while the off diagonal cells have p_2 . In a checkerboard interaction model, Figure 1(d), cells from 4 corners and the center have risk of p_1 , while the other cells have p_2 . The alternative hypotheses generated from two loci are not limited to the above three situations. The number of ways to partition g cells to k groups is given below [8]

$$S(g, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^g.$$

For example, this formula tells us that there are 255 ways to partition the 9 genotypes into 2 groups, 3025 ways to partition 9 genotypes into 3 groups, and so on. Totally, there are 21,146 ways to partition the genotypes into 2 to 9 groups. When there are m candidate loci involved, these 21,146 evaluations will need to be performed for each of $m(m-1)/2$ ways to select two

loci from the candidates. For 10 candidates, there are 45 pairs of loci to be evaluated, and resulting in 951,570 two locus model evaluations. The gene/environmental interaction has the similar formulation.

An ideal statistical approach needs to have the following characters.

1. For a gene/gene or gene/environmental pair, it should possess high power to test its interaction effect on the risk (hypothesis test).
2. Among cells defined by a gene/gene or gene/environmental pair, it should be able to obtain the optimal partition (clustering).
3. If there are n candidate genetic or environmental variables, it should be able to search for all the significant combinations (model search).

In two cancer pharmacogenetics studies introduced in section 3, polymorphisms in multiple genes were genotyped to predict either cancer patient response (survival) or development of side-effects (hot-flashes). We are interested in not only whether these polymorphisms or their combinations are associated with clinical outcomes, but also which combinations have high/low survival or hot-flash risks.

1.3 Current Statistical Methods

Many statistical approaches have been developed to conduct gene/environmental interaction association analyses. They differ in their strategies in searching for optimal models in the space spanned by the genetic/environmental variables, test statistics in inferring the significance, and grouping or clustering capabilities. Table 1 presents a collection of them. More details are introduced below.

Data Structure—Let us assume that the data is composed of m genetic and environmental variables $\{x_1, \dots, x_m\}$, and a binary outcome variable y , in which x s are categorical variables. The questions include whether and what combinations of x s predict y ?

Logistic Regression-Overall Test—Any gene/gene or gene/environmental combination can be tested with an overall chi-square test in a logistic regression. For instance, a two-locus (A/a and B/b) genotype is composed of nine cells defined by combinations of genotypes. A chi-square test of whether these nine genotype cells share the same risk distribution has 8 degrees of freedom. The type-I error is controlled at the per-comparison level. This test searches for all interaction pairs. It can't perform clustering.

Logic Regression - Cross-Validation—Logic regression [9] is an adaptive regression methodology that attempts to construct predictors as Boolean combinations of variables (logic trees). A Boolean combination of a pair of variables represents their interaction. A model is composed of one or multiple logic trees. A step-wise greedy search is used for best model search, and a 10 fold cross-validation is recommended. A permutation test is developed to test the significance of any variables in predicting the outcome. The predicting accuracy is chosen as the test-statistic, the null hypothesis is that all the genetic or environmental variables are null predictors, and the type-I error is controlled at the family-wise level. This data analysis strategy can't perform clustering or partition.

Logic Regression - Monte Carlo Markov Chain (MCMC)—Recently, logic regression is integrated with MCMC search algorithms [10]. It generates many logic models, instead of a single one. As it employs a prior distribution to control the model size, cross-validation is no longer used. At the end of fitting, co-occurrence relative frequencies of variable-pairs are ranked for top interaction selections. This Bayes approach doesn't perform hypothesis test and

clustering. Its MCMC search algorithm outperforms previous step-wise greedy search algorithm.

Multi-Dimension Reduction (MDR)—MDR is a nonparametric approach that tests the existence of high/low risk groups among cells defined by a gene/gene or gene/environmental combinations. It was initially developed for a case/control study. A 10-fold cross-validation strategy is recommended. In the training-set and validation set, cells are classified to high or low risk groups based on a predefined risk threshold. In a case-control study, the threshold is usually defined as 50%. Then the consistent risk group classifications between training and validation sets are treated as the prediction accuracy, and it serves as the test statistic for the interaction. The test statistic empirical distribution is generated through permutation, and it is controlled at the per-interaction level. This MDR approach has the clustering capability, though it only categorized the cells into high and low risk groups [11]. This test searches for all interaction pairs.

1.4 Our Proposed Mixture Model

It appears that only MDR can conduct all hypothesis test, model search, and clustering among current methods. MDR's limitation is that it can only categorize the interaction cells into high/low risk groups, though there may exist more different risk groups. In this paper, we propose a mixture model to address this gene/gene or gene/environmental interaction problem for a binary risk phenotype. Using the two-locus genotype (A/a and B/b) as an example, our goal is to test whether nine cells share the same risk; if they do not, we endeavor to find the optimal genotype partition for different risk groups. The optimal genotype partition is determined by an E-M algorithm in the model's estimation process. A likelihood ratio test (LRT) among multiple mixture distributions for nine cells indicates the significance of the optimal partition. The mixture model and the other methods are introduced in the methods section. The application of our method to two cancer pharmacogenetic studies, and the statistical performances of our mixture model and other methods are extensively investigated and compared in the result section. The summary is presented in the conclusions section.

2. Methods

2.1 Mixture Model

Model Specification—The history of the mixture model's application in genetics can be traced back as far as the 1800s [12], and its important contributions in population genetics have well been documented by Schork [13]. We would like to emphasize that traditional mixture model approaches seek to infer whether a phenotype from the population (such as blood pressure or drug response) is composed of multiple sub-populations that are characterized by possible underlying, unknown genotypes. The mixture model formulation and its estimation procedure is detailed in the book [14].

When genotypes are measured, we propose a mixture model to test genotype/phenotype association and conduct genotype clustering. Unlike a traditional mixture model application [13], genotype cells are grouped in our method instead of individual samples.

In each gene/gene or gene/environmental interaction, let us assume there are G cells, and every cell has n_g ($g=1, \dots, G$) samples. The G cells can be derived from either a gene/gene or a gene/environmental combination. Denote $\mathbf{y}_g = (y_{g1}, \dots, y_{gn_g})$ as the samples in cell g , y_{gi} is a Bernoulli random variable, and $y_{g*} = \sum_{i=1, \dots, n_g} y_{gi}$ is a binomial random variable. If in all cells, $\{\mathbf{y}_g\}_{g=1, 2, \dots, G}$ follow the same distribution (under the null hypothesis), the likelihood function is defined in (1):

$$\Pr(\mathbf{y}|\mathbf{p}_0) = \prod_{g=1}^G \Pr(y_{g\bullet}|\mathbf{p}_0) = \prod_{g=1}^G p_0^{y_{g\bullet}} (1 - p_0)^{n_g - y_{g\bullet}}. \quad (1)$$

If $y_{g\bullet}$ follows one of K different binomial distributions, the complete likelihood function becomes

$$\begin{aligned} \Pr(\mathbf{y}, \mathbf{z}|\mathbf{p}, \mathbf{q}) &= \prod_{g=1}^G \prod_{k=1}^K \Pr^{z_{gk}}(y_{g\bullet}|\mathbf{p}_k) q_k^{z_{gk}}, \\ \Pr(y_{g\bullet}|\mathbf{p}_k) &= \binom{n_g}{y_{g\bullet}} p_k^{y_{g\bullet}} (1 - p_k)^{n_g - y_{g\bullet}}, \quad k=1, \dots, K, \end{aligned} \quad (2)$$

where $(z_{g1}, \dots, z_{gK})$ is a multinomial random vector, $P(z_{gk} = 1) = q_k$. This indicates whether cell g follows the binomial distribution $\Pr(y_{g\bullet} | \mathbf{p}_k)$. Mixture model (2)'s log-likelihood is

$$l(\mathbf{p}, \mathbf{q}; \mathbf{y}, \mathbf{z}) = \sum_{g=1}^G \sum_{k=1}^K z_{gk} [\log \Pr(y_{g\bullet} | \mathbf{p}_k) + \log q_k] \quad (3)$$

Under the null hypothesis, the true model has only one distribution. If we fit the data by a mixture of K -components, any $(p_1 = \dots = p_K = p_0 = p_{\text{true}}, \quad 0 \leq q_k = 1; \sum_{k=1}^K q_k = 1)$ will all achieve the maximum in (3). This problem causes not only numerical difficulty in the mixture model estimation process [15], but also theoretical difficulties in likelihood ratio tests [16,17].

The identification problem was solved in [18] by adding a penalty term into the log-likelihood function (3), by which the penalized likelihood function $pl(\cdot)$, (4), forces $q=0.5$ when it reaches the maximum.

$$pl(\mathbf{p}, \mathbf{q}; \mathbf{y}, \mathbf{z}) = l(\mathbf{p}, \mathbf{q}; \mathbf{y}, \mathbf{z}) + \sum_{k=1}^K \log q_k \quad (4)$$

Our aim is cell clustering, while the work of [18] groups individual observations. This difference leads to distinctive estimation algorithms and asymptotic LRT.

E-M Algorithm—In the estimation step (E-step), the random variable z_{gk} (unobserved) is estimated by equation (5),

$$\hat{z}_{gk} = \frac{\hat{q}_k \Pr(y_{g\bullet} | \hat{\mathbf{p}}_k^{\text{curr}})}{\sum_{l=1}^K \hat{q}_l \Pr(y_{g\bullet} | \hat{\mathbf{p}}_l^{\text{curr}})} \quad (5)$$

As a multinomial random variable, z_{gk} indicates which distribution cell g belongs to. Its estimate by (5) is a probability, and cell g 's distribution or group assignment will be the one with the larger probability:

$$group(g) = \operatorname{argmax}_k \{z_{gk}\}_{k=1,\dots,K}. \quad (6)$$

In the maximization step (M-step),

$$\hat{p}_k = \frac{\sum_{g=1}^G \hat{z}_{gk}^{curr} y_{g\bullet}}{\sum_{g=1}^G \hat{z}_{gk}^{curr} n_g}, \quad \hat{q}_k = \frac{1 + \sum_{g=1}^G \hat{z}_{gk}^{curr}}{K + G} \quad (7)$$

The E- and M-steps are iteratively conducted, and the convergence is monitored based on the relative difference of the penalized likelihood function (4).

Interpretation of the Results—Our null hypothesis is whether G cells share the same binomial distribution. In Figure 1 (a), it displays a null hypothesis for a two loci combination. The alternative hypothesis is that G cells don't share the same binomial distribution, but it may have different appearances. Figure 1(b) presents a situation that the difference in phenotype only depends on locus B/b, while Figure 1(c) and (d) demonstrate that the phenotype difference depends on both two loci. Figure 1(b) is a main effect, and Figure 1(c) and (d) represent interactions. We need to point out that the following likelihood ratio test between 1- and 2-component mixture models can't distinguish these two alternative hypotheses. However, it is the E-M algorithm based clustering procedure and grouping assignment (i.e. equation (5) to (7)), provide the structure of the clustering, hence the interpretation of an alternative hypothesis.

Log-likelihood Ratio Test between 1- and 2-Component Mixtures—To test whether all G cells have the same binomial distribution, or they follow two different distributions, a likelihood ratio test is conducted. The marginal penalized log-likelihoods for a mixture model of two components and one component are listed in (8) and (9)

$$pl_2 = \sum_{g=1,\dots,G} \log \binom{n_g}{y_{g\bullet}} + \sum_{g=1}^G \log [qp_1^{y_{g\bullet}}(1-p_1)^{n_g-y_{g\bullet}} + (1-q)p_2^{y_{g\bullet}}(1-p_2)^{n_g-y_{g\bullet}}] + \log [q(1-q)], \quad (8)$$

$$pl_1 = \log \binom{n}{y_{\bullet\bullet}} + y_{\bullet\bullet} \log p_0 + (n - y_{\bullet\bullet}) \log (1 - p_0) + 2 \log [0.5],$$

$$y_{\bullet\bullet} = \sum_{g=1,\dots,G} y_{g\bullet}, \quad n = \sum_{g=1,\dots,G} n_g. \quad (9)$$

$$\text{Likelihood Ratio Test: } \lambda = -2(pl_2 - pl_1). \quad (10)$$

From the Central Limit Theorem, under the null hypothesis, (8) and (9) can be reformulated as

$$pl_2 = \sum_{g=1}^G \log \left[q\phi \left(\frac{y_{g\bullet} - n_g p_1}{\sqrt{n_g p_1 (1 - p_1)}} \right) + (1 - q)\phi \left(\frac{y_{g\bullet} - n_g p_2}{\sqrt{n_g p_2 (1 - p_2)}} \right) \right] + \log [q(1 - q)] + o_p(1), \quad (11)$$

$$pl_1 = \phi \left(\frac{y_{\bullet\bullet} - np_0}{\sqrt{np_0(1-p_0)}} \right) + 2\log[0.5] + o_p(1), \quad (12)$$

where $\phi(\bullet)$ is the standard normal distribution's density function. Therefore, the likelihood ratio test under the null hypothesis has the asymptotically equivalent distribution to test the hypothesis based on G normal variates $y_{g\bullet} \sim N[n_gp_0, n_gp_0(1-p_0)]$, where $g = 1, 2, \dots, G$. In the supplemental material 1, the location and scale invariance of the test statistics was shown, and it further simplifies the test statistic. The LRT, λ in (10), is equivalent to the hypothesis test based on G standard normal variates. Specifically, we showed that λ is asymptotically equivalent to

$$\begin{aligned} \lambda^* = & -2 \left\{ \max_{\theta_1, \theta_2, q} \sum_{g=1}^G \log [q\phi(w_g - \phi_1) + (1-q)\phi(w_g - \theta_2)] \right. \\ & \left. + \log[q(1-q)] - \max_{\theta_0} \sum_{g=1}^G \log [\phi(w_g - \theta_0)] + 2\log[0.5] \right\}, \end{aligned} \quad (13)$$

where w_g is an i.i.d. standard normal variable. This representation of λ by λ^* has the advantage of not depending on sample size in each genotype cell. The distribution of λ^* depends only on G , the cell number. The value of λ^* can be calculated by first drawing G standard normal variables and then evaluating the expression through E-M algorithm. An empirical distribution of each G is constructed based on 50,000 calculated λ^* . In the supplemental material 2, empirical percentiles of λ^* distributions are reported for $G=2, \dots, 81$. They can thus be utilized to guide p-value calculation.

Log-likelihood Ratio Test between K- and (K+1)-Component Mixtures (K>1)—To test whether G cells follow either K - or $(K+1)$ -component mixture models ($k > 1$), a similar likelihood ratio test is conducted.

$$\begin{aligned} pl_K = & \sum_{g=1}^G \log \left(\frac{n_g}{y_{g\bullet}} \right) + \sum_{g=1}^G \log \left[\sum_{k=1}^K q_k p_k^{y_{g\bullet}} (1-p_k)^{n_g - y_{g\bullet}} \right] + \sum_{k=1}^K \log q_k, \\ & \sum_{k=1}^K q_k = 1, \end{aligned} \quad (14)$$

$$\lambda = -2(pl_{K+1} - pl_K). \quad (15)$$

So far, there is no theoretical guidance for this LRT's asymptotical distribution formulation. Instead, a parametric bootstrap scheme is developed to calculate the p-value. In brief, the data is fitted into a K -component mixture model. Then a simulated data set is generated with this model. This simulated data set is fitted into both K - and $(K+1)$ -component mixture models, and the LRT (15) is calculated. After 10,000 replicated simulation procedures, a null LRT empirical distribution and an empirical p-value are obtained.

2.2 Model Performance Comparison Criteria

Model Translations among Multiple Methods—Although mixture model, MDR, logistic regression, and logic regression possess different model formulations, they can be translated to each other uniquely. Table 2 presents their relationships by using two alternative hypotheses presented in Figure 1 (b) and (c).

Per-comparison Type I Error—For per-comparison type I error, the null hypothesis is whether there is any difference in binary phenotype distribution among cells defined by a gene/gene or gene/environmental interaction. Logistic regression overall test uses chi-square distribution to calculate the per-comparison p-value. Mixture model based likelihood ratio test between 1- and 2-component uses our p-value table (supplemental file 2), which is generated from permutation based empirical distributions (supplemental file 1). MDR also utilizes a permutation test, and the test-statistic is the cross-validation prediction accuracy. In MDR's permutation test, the samples of the binary outcome variable are permuted.

Family-Wise Type I Error—Logic regression usually doesn't have per-comparison type I error control, unless there are only two variables. Given the family-wise type I error, the null hypothesis is that all of the variables are null. In all of the permutation tests, the samples of the binary outcome variable are permuted. In logistic regression, the empirical distribution of maximum of chi-square statistic among $m(m-1)$ interaction pairs are generated through permutations; in both logic regression and MDR, the maximum prediction accuracy empirical distribution under the null is generated; in mixture model, the maximum LR test statistic empirical distribution is generated. Please note that, in our two data analyses, the family-wise type I errors are calculated from these empirical distributions of maximum test statistics.

Powers—In the power calculation, the empirical null distributions of test statistics are generated through parametric bootstraps, in other words, the simulated data sets are generated through null models. If the test-statistic is only controlled at the per-comparison type-I error level, then its empirical null distribution is generated for itself; if the test-statistic is controlled at the family-wise type-I error level, its empirical null distribution is based on the maximum test-statistic.

Three different power calculations are performed for mixture model, MDR, logistic regression, and logic regression with cross-validation.

1. Comparing to the family-wise type I error thresholds, the maximum interaction statistics are calculated to either reject or accept the null hypothesis, i.e. all the genetic and environmental variables are null. This power is call overall power.
2. Comparing to the family-wise or per-comparison type I error thresholds, the interaction statistics are calculated to either reject or accept the null hypothesis for each interaction pair. Then average power is calculated among all non-null interaction pairs.
3. Comparing to the prescribed family-wise or per-comparison type I error thresholds, the interaction analysis statistics are calculated to either reject or accept the null hypothesis for each interaction pair. Then interaction-specific power is calculated for a non-null interaction pair.

False Positive Number and Matched Power—Based on a pre-specified test statistic threshold, the number of false positive interactions can be calculated for each method. At the mean time, various methods' test thresholds can be calibrated such that every method shares the same number of false positive number, hence, different methods can be compared by the matched power. The match power is referred to the average power for non-null interactions. This approach can compare Monte Carlo logic regression to the other ones, because Monte Carlo logic regression is a Bayes approach.

Recovery Probability—Recovery probability is designed to assess how well a method can recover the true model or true clusters. Based on translations among different models, for

example Table 2, their recovery rates can be evaluated and compared, except for Monte Carlo logic regression.

Prediction Accuracy—Prediction accuracy is to assess how well a method can predict the outcome. For a binary outcome, if the predicted probability is above 0.5, then it is classified as 1, else it is 0. The percentage of agreement between predicted and observed variables is expressed as the prediction accuracy. For each real data set and each method, its highest prediction accuracy is chosen from many gene/gene or gene/environmental interactions. Please note that logic regression and MDR use cross-validation to select the best model, while mixture model and logistic regression didn't. Therefore, a fair comparison is the prediction accuracy (fitting) to the whole set with their best models. In a real data set, the prediction accuracy reflects only the fitting. In simulation data sets, we report both fitting and prediction accuracies, where fitting accuracy is calculated from the simulated data sets that a model is trained with, while the prediction is calculated from the data set that a model wasn't trained. Only Monte Carlo logic regression can't perform prediction, since it only has frequencies of predictors appearing in different models.

Computational Implementation—Binomial mixture model program is developed in R (www.r-project.org), a statistics freeware. Logistic regression and logic regression also have well established packages in R. MDR analysis is conducted with its published software, (<http://www.epistasis.org/software.html>).

3. Results

Pharmacogenetic Study of Docetaxel Effects on Hormone Refractory Prostate Cancer Patient's Survival

The Hoosier Oncology Group conducted a randomized phase II trial of weekly docetaxel and vinorelbine (DV) with docetaxel and estramustine phosphate (DE) as a parallel reference arm. The pharmacogenetic component of this project was to test whether a list of selected host genes that play roles in chemotherapy drug transport and metabolism, predict the overall survival. The polymorphisms investigated are summarized in Table 2, and include multi-drug resistance gene ABCB1, and ABCG2, drug degradation genes CYP3A4 and CYP3A5, and genes encoding microtubule associated proteins, MAP4 and MAPT. The clinical endpoint (phenotype) was an indicator variable for patient survival (>15 months or <15 months). An interim analysis of 51 patients was performed using the mixture model, with a goal of searching, grouping, and ranking genetic polymorphisms, and providing priorities for further genotyping.

Genetic polymorphisms and their associations with patient survival were analyzed. Table 4 presents data analyses from mixture model, MDR, and logistic regressions. Per-comparison p-values are reported. In the univariate analysis, only MDR claims ABCG2 significantly predicts patient survival (per-comparison $p = 0.03$). No other methods select any genotypes. In the two-way analysis, both mixture model and logistic regressions select ABCB1/ABCG2, in which per-comparison p-value ranges from 0.02 to 0.04. MDR doesn't select anyone. None of these results are significant at family-wise level. Table 5 presents logic regression's result. In the regular logic regression with 10-fold cross-validation, the optimal model has only ABCG2, and the family-wise p-value for the testing set accuracy is 0.12. In the Monte Carlo logic regression, ABCG2 and ABCB1's selection frequencies rank top 1 and 2 respectively, and their combination ranks 13th.

It appears that two SNPs (ABCB1 3435 and ABCG2 421) are the most significant predictors of patient survival, and not all of the methods agree well with each other. Although ABCB1 3435 and ABCG2 421 are not significant in the mixture model, with p-values of 0.62 and 0.06 respectively, their interaction possesses the strongest evidence (per comparison p-value =

0.02). A more detailed analysis (Figure 2 and Table 6) suggests that the genotype subgroup (ABCB1 TT and ABCG2 CA) has a much higher survival rate ($4/4 = 100\%$) than the other four genotype subgroups ($12/46 = 26\%$). From the mixture model based cluster analysis (Table 6), the subgroup (ABCB1 TT and ABCG2 CA) is then classified as a low risk group, and all the other four subgroups are classified as high risk. The clustering is based on their probability assignments. There is no evidence that there are three risk groups in the two-way analysis, as the p-value for testing two vs three clusters is 0.27. Table 7 presents the optimal prediction accuracies from differential methods. It appears that MDR has slightly lower prediction accuracy (72%) than the others (76%). Please note that this prediction accuracy only reflects the fitting.

Previous studies of the ABCB1 associated P-glycoprotein mediated drug efflux pump have shown minimal, if any, effect on docetaxel plasma concentrations when an inhibitor of this pump is administered [19]. The effect of the ABCG2 polymorphism on docetaxel pharmacokinetics is unknown. The increased survival seen in individuals who carry the (ABCB1 TT and ABCG2 CA) polymorphism may suggest a less functional drug efflux pump leading to increased intracellular docetaxel concentrations and improved cytotoxic activity.

Logistic regression selects the same main and interaction effects (Table 4) as those of the mixture model. However, logistic regression can't perform clustering. MDR doesn't select any significant interaction effects. The optimal models generated by different methods have very comparable prediction accuracies. This example strongly demonstrates the importance of a mixture model in detecting gene-gene interaction, even though individual genes' effects may not be strong. This interim data analysis identified ABCB1 and ABCG2 polymorphisms as priority targets for further investigation.

Pharmacogenetic Study of Tamoxifen Effect on Hot Flashes in Patients with Breast Cancer

A prospective observational study was conducted between 2000 and 2005 in 3 comprehensive cancer centers in the United States for women who were starting standard tamoxifen therapy. Important goals of the study included testing whether estrogen receptor genotypes (ESR1 and ESR2) can predict hot flash development, and whether the genetic effects depend on prior chemo-therapy or menopausal status. The clinical endpoint (phenotype) is the indicator variable for hot flash development after 1-year tamoxifen treatment. Among 300 recruited patients, 124 patients had hot-flash free at the baseline. These 124 patients' genetic and environmental data are summarized in Table 8. Please note that there are some missing data in menopause and genotypes. The following data analyses are based on complete data, and we assume that missing mechanism is random.

Table 9 presents the hot-flash data analysis. Per-comparison p-values are reported for mixture model, MDR, and logistic regression. In the one-way analysis, mixture model and logistic regressions agree very well in selecting ESR2-2, while MDR doesn't select any. In the two-way analysis, mixture model and logistic regression select menopause/ESR2-2, chemo/ESR2-2, ESR1-PvuII/ESR2-2, and ESR1-XbaI/ESR2-2. They all have very comparable significant p-values at both per-comparison and family-wise levels. None is selected by MDR. Logic regression with cross-validation selects ESR2-2, menopausal and chemo, and its family-wise p-value is 0.02. In Monte Carlo logic regression, ESR2-2 and menopause's univariate frequencies rank top three, and their two-way frequency is number 1.

The main effect and interaction between ESR2-2 and menopause are reported in Table 10. In one-way analysis, women who carry the ESR2-02 AA genotype were less likely to develop hot flashes ($5/17 = 29\%$) compared to subjects with the AG/GG genotype ($75/98 = 75\%$), per-comparison p-value = 0.0005; women in peri menopausal status ($11/12 = 92\%$) tends to be more likely to develop hot flash than other women ($73/111 = 66\%$), though it is not significant

(per-comparison p-value = 0.11). In the two-way analysis, pre menopause women with ESR2-2 (AG) genotype and post menopause women with ESR2-2 (AG/AA) genotypes are in the low risk group ($14/35 = 40\%$), and all the others are in the high risk group ($65/76 = 85\%$, per comparison p-value = 0.0018). There is no evidence that there are three risk groups in the two-way analysis, as the p-value for testing two vs three clusters is 0.63. Other significant two-way analyses reported in Table 9, chemo/ESR2-2, ESR1-PvuII/ESR2-2, and ESR1-Xba1/ESR2-2 are predominant ESR2-2's main effect based on their mixture model risk group assignment. The optimal models from different methods provide very comparable predictions (i.e. fitting), 72% to 75% (Table 7).

These results are consistent with results from other tamoxifen pharmacogenetic association studies that suggested that women with ESR2-02 AA genotype are less susceptible to tamoxifen-induced pharmacological effects [20].

Logistic regression selects the same main and interaction effects (Table 9) as those of the mixture model. The overall test of logistic regression can't perform clustering. MDR doesn't select any significant main or interaction effects.

Statistical Simulation: Checkboard and Diagonal Epistatic Models

Two epistatic models, checkboard and diagonal, [21] (Figure 3) are employed to compare different approaches. The numbers in the model represent penetrances of genotype cells from the two loci (A/a and B/b). In the following discussion, we refer to cells with high penetrance as the high risk group, and to cells with low penetrance as the low risk group. The allele frequencies (A, a, B, b) are all fixed at 0.50. Both models are selected because they exhibited interaction effects in the absence of any main effects when genotypes are generated according to Hardy-Weinberg proportions.

Using the simulation strategies outlined by Ritchie [21], we generated 1000 simulated data with and without noise. The noise types include: 50% phenocopy (PC); and 50% genetic heterogeneity (GH). Specifically, PC noise is simulated such that 50% of the affected individuals had genotype combinations that are consistent with the low risk in model 1 or 2; while the other 50% were assumed to be affected by unknown random, environmental effects. In simulating heterogeneity, 50% of affected individuals are due to one high risk two-locus combination, while the other 50% are due to the other high risk combinations. In each data set, 200 cases and 200 controls were simulated. Each case or control has 10 SNPs: two of causative loci with epistatic effects following model 1 and 2; and eight random loci with no effect on case status.

Table 11 presents the power and false positive comparisons among different approaches. The family-wise type I error is set as 0.05 for all methods, and all the significant thresholds are based on 10,000 simulation under the null. The overall power P1 and two-SNP interaction specific power are the same for the mixture model, MDR, and logistic regression, because there is only one pair of true interaction. The overall power and interaction specific power of cross-validation logic regression are defined in the Method Section. In order to compare logic regressions and the other approaches, thresholds are calibrated such all method share the same power of selecting the interactive SNP pair for both epistasis models, hence their corresponding false positive numbers can be compared. The summaries are highlighted below.

- Mixture model and overall-test in logistic regression have comparable power, and they are more powerful than pair-wise comparison in logistic regression and MDR in both epistatic models.
- When all methods have matched power to detect the true interact SNPs, Monte Carlo logic regression performs comparably to mixture model and logistic-regression

overall test. These three methods have lower false positive number than the other methods.

- By introducing the noises (PC and GH) into the model, the power to detect the causal SNP interactions reduces.
- Mixture model, logistic regressions, and logic regressions have slightly higher fitting and prediction accuracies than the MDR has (Table 7).

Statistical Simulation: Higher Order Interactions

A third SNP is introduced to integrate prescribed checkboard model for a three-SNP interaction. This interaction is defined as, if SNP 3 is CC, its penetrance is 0; otherwise, the penetrance follows the checkboard model for SNP 1 and 2 (Figure 3 a). Therefore, 8 of 27 genotype cells among three SNP interaction have a 10% penetrance rate, and the other 19 have 0. Again, phenocopy (PC) percentage is added into the model as a noise. It ranks from 100% to 50%. Only three interacting SNPs are simulated. They all share the same allele frequency of 50%. When PC percentage is 100%, it corresponds to a null hypothesis. In each of 1000 simulated data set, 200 cases and 200 controls are simulated. Family-wise and comparison-wise type I error is equally controlled at 5% level, as one three-way interaction is tested.

The power and recovery probability are reported in table 12. The results suggest that mixture model and logistic regression overall test have slightly higher power than the other methods. At the mean time, mixture model has the best recovery probability to recover the true partition among 27 genotype cells. MDR and logic regression have smaller recovery probabilities, while pair-wise comparison logistic regression can't separate these two clusters at all. Mixture model, logistic regressions, and logic regressions all have comparable fitting and prediction accuracies (Table 7). All the simulations so far point that mixture model and overall test in logistic regression have comparable power.

Statistical Simulation: Logic Model Based Simulations

All the prescribed simulation models favour mixture model. More alternative simulation models, i.e. three logic models, are investigated in this section. For each of these models, 1000 data sets are generated with 1000 individuals and 50 SNP each. Each SNP is in Hardy-Weinberg equilibrium, with a probability of 0.25 of mutant alleles. Linkage disequilibrium between SNPs 1-6 and between SNPs 7-12 by $\text{cor}(X_i, X_{i+1}) = 0.95$ for $i = 1, 2, 3, 4, 5, 7, 8, 9, 10, 11$, and $\text{cor}(X_i, X_{i+1}) = 0$ otherwise. This simulation strategies follows Kooperberg and Ruczinski's work [10]. In stead of using their dominant and recessive SNP codes, we used three dummy variables to code SNPs (see Method Section). The three simulation models are

$$\begin{aligned}\text{Model1:} & \text{logit} [\Pr (Y=1)] = -0.5 + 1.2 \times [(X_{3,11} \vee X_{3,12}) \wedge (X_{9,11} \vee X_{9,12})], \\ \text{Model2:} & \text{logit} [\Pr (Y=1)] = -0.5 + 0.4 \times (X_{13,11} \vee X_{13,12}), \\ \text{Model3:} & \text{logit} [\Pr (Y=1)] = -0.5 + 1.5 \times (X_{16,11} \vee X_{16,12}) + [(X_{16,11} \vee X_{16,12}) \wedge (X_{17,11} \vee X_{17,12})].\end{aligned}$$

Table 7, 13 and 14 present results. The highlights are summarized.

- When the family-wise type I error is controlled at 5% level, mixture model and the overall test in logistic regression have comparable power, and they are much powerful than all the other methods. MDR is the least powerful method (Table 13).
- When all the methods share the matched power derived from the cross-validation logic regression, in detecting main effect of model 2, mixture model, logistic regression, and logic regression have comparable FP, but MDR has a much higher FP; in detecting main and interaction effects of model 3, mixture model and Monte

Carlo logic regression have the smallest FP, and MDR's FP is much higher (Table 13).

- When all the methods share the matched power derived from the cross-validation logic regression, in detecting interaction effect of model 1, mixture model and the overall test in logistic regression have relative high FP than logic regressions. However, the majority of the selected interactions are in the LD region. MDR select much more FP interaction, and the majority of them are not in the LD region (Table 13 and 14).
- Logic regression has only slightly higher fitting and prediction accuracies than the other approaches. (Table 7).

Statistical Simulation: Analysis Workshop in Pharmacogenetics Research Network

A hypothetical disease called “tenureitus” is investigated. The study is designed to include 557 unrelated patients. Among those 557 patients, 206 of them were selected, and denser genotyping is conducted. Hence, around 60% of individuals have missing genotypes. There are totally 188 SNPs typed in three candidate genes: R1SNP1 - R1SNP88, R2SNP1 - R2SNP28, and R3SNP1 - R3SNP74. At the mean time, 12 baseline variables are recorded, which include alcohol, DUI, IVY, degree, gender and etc, in which IVY indicates whether an individual has an Ivy League education background. One hundred simulated data sets are simulated to mimic gene/environmental interactions in this association studies. One of the phenotypes is binary: RIVALSIDE. Its simulation model is as following.

$$\begin{aligned} \text{RIVALSIDE} &= 1 \{x_2 < \text{logit}(p_2)\}, \\ x_2 &\sim U(0, 1); \quad p_2 = 1.05 \times \text{IVY} + 4 \times \text{PENHAP} + 10 \times \text{IVY} \times \text{PENHAP}, \\ \text{PENHAP} &= (0, 1/8, 1) \quad \text{for haplotype } (0, 1, 2) \text{ formed by SNPs, } R2\text{SNP5} - R2\text{SNP10}. \end{aligned}$$

As missing data severely limits the application of logic regressions, we conduct simulation data analysis and power comparisons only among mixture model, MDR, and logistic regression. The type I error rate is controlled at the per comparison level of 5%. Table 15 summarizes the results. In brief all three methods have very comparable power performance in detecting both main and interaction effects. In particular, all three analyses reach 100% power in detect IVY effect, and its interactions with all the other SNPs in the haplotype regions. In addition, all three methods have comparable performances in fitting and predicting the outcome.

Statistical Simulation: Finite or Unequal Sample Size Effects on the Asymptotic Distribution of the Log-likelihood

The LRT in (13) is based on a large sample theory, which is not necessarily true in general. Two simulation studies were conducted to address how sensitive the LRT's asymptotic distribution is to our assumption. In the first simulation, two genotype cells are generated in a simulated case/control study, which have sample sizes (n_1, n_2) . Both genotype cells share the same case frequencies of 0.5. A LRT is calculated to test whether two genotype cells have the same case frequency or not. To obtain the empirical distribution of the LRT statistics, 10,000 data sets are generated. Four sample size combinations are tested: $(n_1, n_2) = (5, 5), (500, 500), (5, 80), \text{ and } (25, 400)$. Their LRT's empirical distributions are plotted against the asymptotic empirical distribution calculated in the method section (QQ-plots in figure 4). The QQ-plots (figure 4) suggest that though LRT's empirical distribution in small sample size (i.e. $(n_1, n_2) = (5, 5)$) is slightly different from its asymptotic distribution, its tail quantile's approximation is very good. Hence it will not affect p-value calculation for a significant result. On the other hand, un-equal sample size has little effect on the asymptotic distribution of the LRT.

In the second simulation, nine genotype cells are generated in a simulated case/control study, which have sample sizes $(n_1, \dots, n_9)$. Both genotype cells share the same case frequencies of 0.5. A LRT is calculated to test whether two genotype cells have the same case frequency or not. To obtain the empirical distribution of the LRT statistics, 10,000 data sets are generated. Four sample size combinations are tested: $(n_1, \dots, n_9) = (5, \dots, 5)$, $(500, \dots, 500)$, $(5, 10, 15, 20, 25, 30, 35, 40, 45)$, and $(100, 100, 100, 100, 500, 500, 500, 500, 500)$. Their LRT's empirical distributions are plotted against the asymptotic empirical distribution calculated in the method section (QQ-plots in figure 5). The QQ-plots suggest that LRT's empirical distribution in small sample size (Figure 5a) is slightly under estimated, which means a conservative p-value calculation based on the asymptotic distribution. On the other hand, un-equal sample size won't affect the tail probability much (Figure 5 c d). The approximation of big sample size (Figure 5 b) works well.

Computational Speed

This mixture model is implemented in *R*. The speed of the E-M algorithm depends on the number of genotype cells. As the cell number ranges from two to nine in regular gene-gene and gene-environmental interactions, the model fitting is usually fast. Specifically, as the p-value calculation is guided by the LRT's asymptotic distribution, its computation time is negligible. Therefore, this mixture model approach has high computational efficiency. For example, it takes only 1 hour to conduct 1,000,000 LRT on a laptop computer (Dell Latitude D600, 2.0GHz processor, and 1.0GB RAM).

4. Conclusions

Our proposed mixture model approach can simultaneously test gene/gene or gene/environmental interactions, and group interaction cells in phenotype prediction. Our theoretical derivation provides guidance for p-value calculation for the LRT test of one-cluster vs two-cluster, though LRT tests on more clusters still rely on the computational expensive permutation scheme.

Extensive simulation studies demonstrate that mixture model and the overall test in logistic regression constantly possess the best power to test whether all the cells among multi-way combinations have the same binomial distribution. MDR sometimes has comparable performance, and sometimes doesn't. It works better in case/control studies. Logic regression, when integrated with the Monte Carlo approach, its performance becomes comparable to mixture model and the overall test in logistic regression. All these methods' performances are evaluated under either per-comparison type I error or family-wise type I error control.

At the mean time, the mixture model approach has the highest recovery probability to recover the true partition in the simulation studies. MDR and cross-validation based logic regression have lower probability to recovery the true partition, because MDR has lower power to detect the true predictors, and the cross-validation based logic regression tends to include more false positive predictors. Pair-wise comparison based logistic regression can't perform clustering at all. All of the methods provide comparable predictions to the outcomes.

The scope of this paper focuses on a candidate gene approach in association studies. In other word, candidate genes have been screened by either biology knowledge or a genome wide scan study. If we start from a genome-wide study, the performances of these discussed methods need more assessments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research is sponsored by NIH grants, R01 GM74217 (LL) U-01 GM61373 (DF) and R-01 GM56898 (DF).

References

- Desai M, Tanus-Santos J, Li L, Gorski JC, Arefayene M, Liu Y, Desta Z, Flockhart D. Pharmacokinetics and QT interval pharmacodynamics of oral haloperidol in poor and extensive metabolizers of CYP2D6. *The Pharmacogenetics Journal* 2003;3:105–113.
- Ho H, Pinto A, Hall SD, Flockhart DA, Li L, Skaar T, Cadman P, O'Connor DT, Wagner U, Fineberg N, Weinberger MH. The Association Between the CYP3A5 Genotype and Blood Pressure. *Hypertension* 2005;45:294–298. [PubMed: 15596575]
- Jin Y, Desta Z, Stearns V, Ward B, Ho H, Lee J, Skaar T, Storniolo AM, Li L, Araba A, Blanchard R, Nguyen A, Ullmer L, Hayden J, Lemler S, Weinshilboum RM, Rae JM, Hayes DF, Flockhart D. Effects of CYP2D6 Genotype and Antidepressants on Tamoxifen Metabolism During Adjuvant Breast Cancer Treatment. *Journal of National Cancer Institute* 2005;97:30–39.
- Borgesm S, Desta Z, Li L, Skaar T, Ward B, Nguyen A, Jin Y, Storniolo M, Nikoloff M, Koch W, Hayes D, Stearns V, Flockhart D. Quantitative effect of CYP2D6 genotype and inhibitors on tamoxifen metabolism: implication for optimization of breast cancer treatment. *Clinical Pharmacology and Therapeutics*. 2006in press
- Rae J, Goetz M, Hayes D, Li L, Stornio A, Stearns V, Flockhart D. Letter to the Editor. *Breast Cancer Research*. 2006in press
- Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. *Ann Med* 2002;34:88–95. [PubMed: 12108579]
- Culverhouse R, Klein T, Shannon W. Detecting Epistatic Interactions Contributing to Quantitative Traits. *Genetic Epidemiology* 2004;27:141–152. [PubMed: 15305330]
- Comtet, L. Advanced combinatorics: the art of infinite expansions. Reidel; Boston: 1974.
- Ruczinski I, Kooperberg C, LeBlanc ML. Logic Regression. *Journal of Computational and Graphical Statistics* 2003;12:475–511.
- Kooperberg C, Ruczinski I. Identifying Interacting SNPs Using Monte Carlo Logic Regression. *Genetic Epidemiology* 2005;28:157–170. [PubMed: 15532037]
- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147. [PubMed: 11404819]
- Pearson KP. Contributions to the mathematical theory of evolution. *Philosophical Transactions* 1895;A186:342–414.
- Schork NJ, Allison DB, Thiel Bonnie. Mixture distributions in human genetics research. *Statistical Methods in Medical Research* 1996;5:155–178. [PubMed: 8817796]
- McLachlan, G.; Peel, D. *Finite Mixture Model*. 2000.
- Seidel W, Mosler K, Alker M. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*. 2000in press
- Hartigan, JA. In: Neyman, J.; Kiefer, J., editors. A failure of likelihood asymptotics for normal mixture; Proceedings of the Berkeley Conference; 1985; p. 807-810.
- Ghosh, JK.; Sen, PK. In: Neyman, J.; Kiefer, J., editors. On the asymptotics performance of the log-likelihood ratio statistics for the mixture model and related results; Proceedings of the Berkeley Conference; 1985; p. 799-806.
- Chen J, Kalbfleisch JD. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of Royal Statistical Society, Series B* 2001;63:19–29.
- Kahlert S, van Eickels M, Vetter H, Meyer R, Grohe C. Estrogen receptor rapidly activates the IGF-1 receptor pathway. *J. Biol Chem* 2000;275:18447–18453. [PubMed: 10749889]N.S.
- Hu Z, Ho H, Skaar T, Rehman MI, Jin Y, Desta Z, et al. Thyroid binding globulin increase following tamoxifen treatment is associated with polymorphisms in the estrogen receptors gene. *Drug Metab Rev* 2004;36(Supplement 1)

21. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimension reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiology* 2003;24:150–157. [PubMed: 12548676]

Null and Alternative Hypotheses

(a) H_0

	BB	Bb	bb
AA	p0	p0	p0
Aa	p0	p0	p0
aa	p0	p0	p0

(b) H_A : main effect

	BB	Bb	bb
AA	p2	p2	p1
Aa	p2	p2	p1
aa	p2	p2	p1

(c) H_A : diagonal interaction

	BB	Bb	bb
AA	p1	p2	p2
Aa	p2	p1	p2
aa	p2	p2	p1

(d) H_A : checkerboard interaction

	BB	Bb	bb
AA	p1	p2	p1
Aa	p2	p1	p2
aa	p1	p2	p1

Figure 1.

Null and Alternative Hypotheses (a) null hypothesis; (b) alternative hypothesis with a main effect; (c) alternative hypothesis with a diagonal gene/gene interaction; and (d) alternative hypothesis with a checkerboard gene/gene interaction.

Figure 2.

Within the genotype subgroups defined by ABCB1 and ABCG2, genotype subgroup (ABCB1 TT and ABCG2 CA) has a much higher survival rate (4/4=100%) than the other four genotype subgroups (12/46 = 26%).

Gene-Gene Interaction Model

Checkboard Model				Diagonal Model			
AA	Aa	aa		AA	Aa	aa	
0.0	0.1	0.0	BB	0.0	0.0	0.1	BB
0.1	0.0	0.1	Bb	0.0	0.05	0.0	Bb
0.0	0.1	0.0	bb	0.1	0.0	0.0	bb

Figure 3.

The numbers in the model represent penetrances of genotype cells from the two loci (A/a and B/b). We refer to cells with high penetrance as the high risk group, and to cells with low penetrance as the low risk group. The allele frequencies (A, a, B, b) are all fixed at 0.50. Both models are selected because they exhibited interaction effects in the absence of any main effects when genotypes are generated according to Hardy-Weinberg proportions.

QQ-plots

Figure 4.

Two cell LRT test for two clusters. The thick lines are QQ-curves. The thin lines are the diagonal straight lines. The y-axis is the empirical quantile from finite samples; and x-axis is empirical quantile from the asymptotic distribution.

QQ-plots

Figure 5.

Nine cell LRT test for two clusters. The thick lines are QQ-curves. The thin lines are the diagonal straight lines. The y-axis is the empirical quantile from finite samples; and x-axis is empirical quantile from the asymptotic distribution. (a) All nine cells have 5 samples; (b) all nine cells have 500 samples; (c) the samples sizes are (5, 10, 15, 20, 25, 30, 35, 40, 45); and (d) the samples are (100, 100, 100, 100, 500, 500, 500, 500).

Table 1

Characters of Different Gene/Gene and Gene/Environmental Interactions Data Analysis Methods

Methods	Hypothesis Test	Model Search	Clustering	Type-I error
Logistic Regression (Overall)	Yes	All Pair-wise	No	Per-Comparison [*]
Logic-Regression (Cross-Validation)	Yes	Step-wise Greedy Search	No	Family-wise ^{**}
Logic Regression (Monte Carlo)	No	MCMC	No	No
Multi-Dimension Reduction (MDR)	Yes	All Pair-wise	Yes	Per-Comparison [*]
Mixture Model	Yes	All Pair-wise	Yes	Per-Comparison [*]

Per- Comparison

* type-I error means the p-value for testing whether there is any difference in phenotype among combination cells defined by A pair of variables. Family-wise

** type I error is defined as the p-value for testing whether there is any difference among combination cells defined by ANY pairs of variables.

Table 2

Model Translations among Multiple Methods

Methods	Main Effect Figure 1 b	Interaction Figure 1 c
Mixture Model	$\Pr(\mathbf{y}, \mathbf{z} \mid p_1, p_2, q) = \prod_{g=1}^G \Pr^{z_g} \left(y_{g \cdot} \mid p_1 \right) q^{z_g} \Pr^{1-z_g} \left(y_{g \cdot} \mid p_2 \right) (1-q)^{1-z_g}$ $group(z_g) = \begin{cases} 1, & g = 3, 6, 9 \\ 2, & \text{otherwise} \end{cases}$	$group(z_g) = \begin{cases} 1, & g = 1, 5, 9 \\ 2, & \text{otherwise} \end{cases}$
MDR	$\Pr(y_{gi}) = \begin{cases} p_1, & g = 3, 6, 9 \\ p_2, & \text{otherwise} \end{cases}$	$\Pr(y_{gi}) = \begin{cases} p_1, & g = 1, 5, 9 \\ p_2, & \text{otherwise} \end{cases}$
Logistic Regression	$\text{logit}[\Pr(y_{gi})] = \beta_0 + \beta_1 x_{gi}, \beta_0 = \log \frac{p_2}{1-p_2}, \beta_1 + \beta_0 = \log \frac{p_1}{1-p_1}$ $x_{gi} = \begin{cases} 1, & g = 3, 6, 9 \\ 0, & \text{otherwise} \end{cases}$	$x_{gi} = \begin{cases} 1, & g = 1, 5, 9 \\ 0, & \text{otherwise} \end{cases}$
Logic Regression	$\text{logit}[\Pr(y_{g_1 g_2 i})] = \beta_0 + \beta_1 L(g_1, g_2)$ $L(g_1, g_2) = (g_2 = 3)$	$L(g_1, g_2) = \begin{bmatrix} (g_1 = 1 \wedge (g_2 = 1)) \\ \vee [(g_1 = 2) \wedge (g_2 = 2)] \\ \vee [(g_1 = 3) \wedge (g_2 = 3)] \end{bmatrix}$

Note: $g = (1, \dots, 9)$ are corresponding to genotype cells (AA/BB, AA/Bb, AA/bb, Aa/BB, Aa/Bb, Aa/bb, aa/BB, aa/Bb, aa/bb) respectively; $g_1 = (1, 2, 3)$ are corresponding to genotypes (AA, Aa, aa) respectively; and $g_2 = (1, 2, 3)$ are corresponding to genotypes (BB, Bb, bb), respectively.

Table 3

Genetic Polymorphisms in Prostate Pharmacogenetic Study

Gene	Polymorphisms		
ABCB1 3435	C/C (10)	C/T (29)	T/T (12)
ABCG2 421	C/C (44)	C/A (6)	A/A (0)
CYP3A4*1B	A/A (43)	A/G (5)	G/G (2)
CYP3A5*3C	A/A (2)	A/G (5)	G/G (44)
MAP4 68	G/G (46)	G/A (5)	A/A (0)
MAP4 1280	C/C (25)	C/A (18)	A/A (5)
MAPT-13	G/G (3)	G/A (18)	A/A (29)
MAPT 3674	A/A (30)	G/A (17)	G/G (3)

Table 4

Prostate Pharmacogenetic Study Data Analysis I Mixture Model, MDR, Logistic Regressions

Effects	Mixture Model	MDR	Logistic Regression overall test
One-way:			
ABCB1	(0.62, 0.99)	(0.52, 0.97)	(0.24, 0.73)
ABCG2	(0.06, 0.25)	(0.03, 0.13)	(0.09, 0.36)
CYP3A4	(0.09, 0.36)	(0.36, 0.88)	(0.10, 0.39)
CYP3A5	(0.09, 0.36)	(0.09, 0.36)	(0.20, 0.65)
MAP4.1	(1.00, 1.00)	(0.58, 0.98)	(0.41, 0.92)
MAP4.2	(1.00, 1.00)	(0.81, 0.99)	(0.34, 0.86)
MAPT.1	(0.55, 0.97)	(0.86, 0.99)	(0.42, 0.92)
MAPT.2	(0.61, 0.98)	(0.33, 0.85)	(0.47, 0.95)
Two-way:			
ABCB1/ABCG2	(0.02, 0.09)	(0.66, 0.99)	(0.04, 0.17)
ABCB1/CYP3A4	(0.41, 0.92)	(0.17, 0.59)	(0.48, 0.95)
ABCB1/CYP3A5	(0.48, 0.95)	(0.59, 0.98)	(0.37, 0.89)
ABCB1/MAP4.1	(1.00, 1.00)	(0.66, 0.99)	(0.31, 0.83)
ABCB1/MAP4.2	(1.00, 1.00)	(0.32, 0.84)	(0.67, 0.99)
ABCB1/MAPT.1	(0.33, 0.85)	(0.59, 0.98)	(0.53, 0.97)
ABCB1/MAPT.2	(0.32, 0.84)	(0.47, 0.95)	(0.52, 0.97)
ABCG2/CYP3A4	(0.09, 0.36)	(0.47, 0.95)	(0.09, 0.36)
ABCG2/CYP3A5	(0.08, 0.32)	(0.37, 0.89)	(0.08, 0.32)
ABCG2/MAP4.1	(0.17, 0.59)	(0.11, 0.42)	(0.22, 0.69)
ABCG2/MAP4.2	(0.16, 0.56)	(0.65, 0.99)	(0.16, 0.56)
ABCG2/MAPT.1	(0.46, 0.94)	(0.65, 0.99)	(0.30, 0.81)
ABCG2/MAPT.2	(0.45, 0.94)	(0.50, 0.96)	(0.44, 0.93)
CYP3A4/CYP3A5	(0.08, 0.32)	(0.31, 0.83)	(0.36, 0.88)
CYP3A4/MAP4.1	(0.43, 0.93)	(0.78, 0.99)	(0.23, 0.71)
CYP3A4/MAP4.2	(0.44, 0.93)	(0.95, 0.99)	(0.46, 0.94)
CYP3A4/MAPT.1	(0.10, 0.39)	(0.32, 0.84)	(0.39, 0.90)
CYP3A4/MAPT.2"	(0.13, 0.48)	(0.32, 0.84)	(0.23, 0.71)
CYP3A5/MAP4.1	(0.43, 0.93)	(0.81, 0.99)	(0.47, 0.95)
CYP3A5/MAP4.2	(0.43, 0.93)	(0.92, 0.99)	(0.41, 0.92)
CYP3A5/MAPT.1	(0.09, 0.36)	(0.38, 0.89)	(0.46, 0.94)
CYP3A5/MAPT.2	(0.12, 0.45)	(0.21, 0.67)	(0.33, 0.85)
MAP4.1/MAP4.2	(0.44, 0.93)	(0.85, 0.99)	(0.33, 0.85)
MAP4.1/MAPT.1	(0.55, 0.97)	(0.11, 0.42)	(0.32, 0.84)
MAP4.1/MAPT.2	(0.50, 0.96)	(0.58, 0.98)	(0.32, 0.84)
MAP4.2/MAPT.1	(0.34, 0.86)	(0.88, 0.99)	(0.48, 0.95)
MAP4.2/MAPT.2	(0.38, 0.89)	(0.68, 0.99)	(0.36, 0.88)
MAPT.1/MAPT.2	(0.50, 0.96)	(0.36, 0.88)	(0.60, 0.98)

Note: (per-comparison p-value, family-wise p-value).

Table 5**Prostate Pharmacogenetic Study Data Analysis II Logic Regressions**

Logic Regression		
CV:		
family-wise p-value = 0.12		
logit(Pr(Survival<15 month)) = 1.85×1{ABCG2 = 0}		
<hr/>		
Monte Carlo Logic Regression		
Top three main effects		
freq.		
<hr/>		
ABCG2=1		0.07388312
ABCG1=0		0.07301971
MAPT1=2		0.05148383
<hr/>		
<hr/>		
Top 7 two-way interactions		
Factor1	Factor 2	Freq.
<hr/>		
ABCB1=1	MAPT.1=2	0.0081160421
ABCG2=0	MAPT.1=2	0.0075239904
ABCB1=2	MAPT.1=2	0.0073759775
ABCG2=1	CYP3A4=2	0.0072773022
ABCG2=1	MAPT.2=0	0.0067592570
ABCB1=1	MAPT.2=0	0.0067345882
ABCG2=0	MAP4.1=0	0.0066852505
ABCG2=1	MAPT.1=2	0.0066605817
ABCG2=0	CYP3A5=2	0.0064878999
ABCG2=1	MAPT.1=1	0.0064138935
ABCG2=1	CYP3A5=2	0.0063892247
ABCG2=0	MAPT.2=1	0.0063645558
ABCB1=0	ABCG2=0	0.0063152182
<hr/>		

Table 6

High/Low Risk Group Assignment Based on the Mixture Model in Prostate Pharmacogenetic Study

ABCB1	ABCG2	Survival<15month frequency	Risk Prediction (z_g)	
			low	high
CC	CC	3/9 = 33%	0.23	0.76
CT	CC	7/27 = 26%	0.12	0.88
TT	CC	2/8 = 25%	0.19	0.81
CT	CA	0/2 = 0%	0.08	0.92
TT	CA	4/4 = 100%	0.99	0.01

Note: mixture model per-comparison p-value: 1 vs 2 (0.02); 2 vs 3 (0.27).

Table 7**Model Based Prediction Accuracies for Clinical Examples and Simulation Studies**

Data Sets	MM*	MDR**	Methods Logistic	Logic
<i>Clinical Examples:</i>				
Prostate PG Trial	76%	72%	76%	76%
Tamoxifen Trial	75%	75%	75%	72%
<i>Simulation Studies:</i>				
<i>Fitting Accuracies:</i>				
Checkboard Model	78%	71%	79%	74%
Diagonal Model	77%	72%	77%	75%
Three-Way Interaction	76%	75%	76%	75%
<i>Logic Models</i>				
Model1	83%	80%	82%	85%
Model2	72%	73%	72%	76%
Model3	79%	79%	79%	82%
Model1 (LD SNPs)	85%	82%	84%	87%
GWA Data	83%	83%	84%	N/A
<i>Simulation Studies:</i>				
<i>Predicting Accuracies:</i>				
Checkboard Model	73%	69%	73%	72%
Diagonal Model	73%	72%	77%	71%
Three-Way Interaction	74%	75%	74%	75%
<i>Logic Models</i>				
Model1	80%	80%	80%	83%
Model2	69%	71%	70%	74%
Model3	74%	75%	72%	74%
Model1 (LD SNPs)	82%	82%	82%	83%
GWA Data	82%	81%	82%	N/A

Note:

Both checkboard, diagonal model, and three-way interaction use 50% phenocopy (PC).

MM* mixture model

MDR** multi-dimension reduction

Table 8

Genetic and Environment Variables within the Tamoxifen Trial

Variable	Categories		
Menopause	pre (56)	peri (12)	post (55)
Prior Chemo	no (73)	yes (51)	
ESR2-2	G/G (57)	A/G(41)	A/A(17)
ESR1-PvuII	C/C(24)	C/T(59)	T/T(36)
ESR1-XbaI	G/G(11)	G/A(51)	A/A(51)

Table 9

Hot-flash Data Analysis

Effects	Mixture Model (per-comparison p-value, family-wise p-value)	MDR	Logistic Regre overall test
<hr/>			
One-way:			
Menopause	(0.11, 0.35)	(0.39, 0.84)	(0.08, 0.27)
Chemo	(0.85, 0.99)	(0.14, 0.43)	(0.33, 0.78)
ESR2-2	(0.0005, 0.001)	(0.10, 0.32)	(0.0004, 0.001)
ESR1PvuII	(0.43, 0.88)	(0.56, 0.95)	(0.81, 0.99)
ESR1Xba	(0.33, 0.78)	(0.95, 0.99)	(0.91, 0.99)
<hr/>			
Two-way:			
Menopause/Chemo	(1.00, 1.00)	(0.14, 0.43)	(0.28, 0.71)
Menopause/ESR2-2	(0.0018, 0.006)	(0.06, 0.21)	(0.0006, 0.002)
Menopause/ESR1PvuII	(1.00, 1.00)	(0.67, 0.98)	(0.58, 0.96)
Menopause/ESR1Xba	(1.00, 1.00)	(0.37, 0.83)	(0.70, 0.98)
Chemo/ESR2-2	(0.004, 0.015)	(0.37, 0.82)	(0.0006, 0.002)
Chemo/ESR1PvuII	(0.56, 0.95)	(0.37, 0.82)	(0.62, 0.97)
Chemo/ESR1Xba	(0.59, 0.96)	(0.34, 0.79)	(0.67, 0.98)
ESR2-2/ESR1PvuII	(0.019, 0.07)	(0.34, 0.79)	(0.027, 0.098)
ESR2-2/ESR1Xba	(0.016, 0.059)	(0.32, 0.77)	(0.031, 0.112)
ESR1PvuII/ESR1Xba	(1.000, 1.000)	(0.88, 0.99)	(0.39, 0.85)
<hr/>			
Logic Regression (CV)			
logit(Pr(Hotflash = 1)) =			
3.46×(^1{ESR2_2=AA}∨^1{Menopause = Pre}^1{Chemo=Yes}))			
family-wise p-value = 0.02			
<hr/>			
<hr/>			
Monte Carlo Logic Regression			
Top three main effects		freq.	
1(ESR2-2=AA)		0.555	
1(Meno=Peri)		0.069	
1(ESR2-2=GG)		0.05634814	
<hr/>			
<hr/>			
Top 5 two-way interactions			
Factor1	Factor2	Freq.	
<hr/>			
1(Menopause=Peri)	1(ESR2-2 = AA)	0.0597	
1(ESR2-2=AA)	1(ESR2-2 = GG)	0.0377	
1(Chemo=Yes)	1(ESR2-2 = AA)	0.0368	
1(ESR2-2=AA)	1(ESR2-2 = AG)	0.0316	
1(Menopause=Pre)	1(ESR2-2 = AA)	0.0255	

Table 10**High/Low Risk Group Assignment Based on the Mixture Model**

Hot-flash Frequency		Hot-flash Risk Prediction (z_g)	
		low	high
ESR2-2			
AA	5/17=29%	1.00	0.00
AG	31/41=76%	0.00	1.00
GG	44/57=77%	0.00	1.00
Mixture model p-value: 1 vs 2 (0.00005)			
Menopause			
Pre	40/56 = 71%	1.00	0.00
Peri	11/12 = 92 %	0.15	0.85
Post	33/55 = 60%	1.00	0.00
Mixture model p-value: 1 vs 2 (0.11)			
Menopause			
Pre	ESR2-2		
Pre	AA	1/8 = 12.5%	1.00
Pre	AG	14/15 = 93.0%	0.01
Pre	GG	22/29 = 75.9%	0.13
Peri	AA	1/1 =100.0%	0.41
Peri	AG	7/7 =100.0%	0.05
Peri	GG	3/4 = 75.0%	0.46
Post	AA	3/8 = 37.5%	0.98
Post	AG	10/19 = 52.6%	1.00
Post	GG	18/23 = 78.3%	0.09
Mixture model p-value: 1 vs 2 (0.0018); 2 vs 3 (0.67).			

Table 11

Comparisons for Checkboard and Diagonal Epistasis Models

Source of Noise	Checkboard Model (P1 [*] , P2 ⁺ , P3 [#] , FP ^{\$})	Diagonal Model (P1 [*] , P2 ⁺ , P3 [#] , FP ^{\$})
Mixture Model		
None	(100%, 100%, 100%, 0.00)	(100%, 100%, 100%, 0.00)
GH	(100%, 100%, 100%, 0.00)	(100%, 100%, 100%, 0.00)
PC	(88%, 88%, 64%, 0.00)	(99%, 99%, 70%, 0.00)
MDR		
None	(100%, 100%, 100%, 0.00)	(100%, 100%, 100%, 0.00)
GH	(98%, 98%, 100%, 0.28)	(100%, 100%, 100%, 0.06)
PC	(47%, 45%, 64%, 0.19)	(65%, 63%, 70%, 0.09)
Logistic Regression: overall test.		
None	(100%, 100%, 100%, 0.00)	(100%, 100%, 100%, 0.00)
GH	(100%, 100%, 100%, 0.00)	(100%, 100%, 100%, 0.00)
PC	(90%, 90%, 64%, 0.02)	(99%, 99%, 70%, 0.00)
Logic Regression: cross-validation		
None	(100%, 99%, 99%, 2.17)	(100%, 100%, 100%, 2.03)
GH	(100%, 100%, 100%, 1.72)	(100%, 100%, 100%, 1.36)
PC	(32%, 29%, 64%, 0.84)	(39%, 37%, 70%, 0.89)
Logic Regression: Monte Carlo		
None	(NA, NA, 99%, 0.00)	(NA, NA, 100%, 0.00)
GH	(NA, NA, 100%, 0.00)	(NA, NA, 100%, 0.00)
PC	(NA, NA, 64%, 0.03)	(NA, NA, 70%, 0.01)

* P1 is the overall power under 5% family-wise type I error

⁺ P2 is the power to detect two causal SNP's epistasis model under 5% family-wise type I error

[#] P3 is the matched power of selecting two causal SNP with cross-validation in logic regression

^{\$} FP is the number of selected false positive SNPs, when all the methods share the matched power P3.

Table 12

Power and Recovery Percentile for Three Gene Interaction

Methods	100%	90%	PC Percentile 70%	50%
<i>Power:</i>				
Mixture Model	5%	10.3%	67.0%	100.0%
MDR	5%	07.7%	43.7%	96.9%
Logistic Regression overall test	5%	09.6%	65.7%	99.9%
Logic Regression (cross-validation)	5%	07.1%	58.8%	99.8%
<i>Recovery Probability:</i>				
Mixture Model	N/A	5.3%	62.0%	99.8%
MDR	N/A	4.8%	37.5%	95.3%
Logistic Regression overall test	N/A	N/A	N/A	N/A
Logic Regression (cross-validation)	N/A	3.5%	43.3%	92.5%

Note: family-wise and comparison-wise type I error is equally controlled at 5% level, as one three-way interaction is tested.

Table 13

Power and False Positive Comparisons for Logic Model Based Simulations

Methods	Model1	Model2	Model3
[*] <i>Power</i>			
Mixture Model	0.83	0.11	0.92
MDR	0.06	0.04	0.10
Logistic Regression overall test	0.86	0.09	0.97
Logic Regression - cross-validation	0.110	0.06	0.11
- Monte Carlo	N/A	N/A	N/A
⁺ <i>FP</i>			
[#] <i>Matched Power</i>			
Mixture Model	0.14	0.33	0.22
MDR	1.94	0.864	0.000
Logistic Regression overall test	19.02	14.92	10.55
Logic Regression - cross-validation	2.048	0.718	0.002
- Monte Carlo	0.906	0.842	0.48
	0.506	0.946	0.002

^{*} *Power* is the main or interaction effects' power under each model, when the family wise type I error is control at 5% level.

[#] *Matched Power* refers to the cross-validation logic regression's power in selecting main or interaction effects.

⁺ *FP* is the number of false positive with matched power.

Table 14

Interactions Selected from the LD Region in Model 1

	Total	(3, 9)	LD Interaction	LD%
Mixture Model	2.08	0.144	2.018	97%
MDR	19.172	0.144	3.540	18.5%
Logistic Regression				
overall test	2.186	0.144	2.124	0.97%
Logic Regression				
- cross-validation	1.050	0.144	0.724	67%
- Monte Carlo	0.652	0.144	0.612	94%

Note: LD interaction means the interaction SNPs are in the simulated LD region of SNP 3 and 9. All the methods share the matched power derived from the cross-validation logic regression.

Table 15

Power and Frequencies Comparison for the GAW Data

Effect	Mixture Model	MDR	Logistic (overall test)
Main Effect:			
IVY	100%	100%	100%
R2SNP5	12%	7%	12%
R2SNP6	8%	8%	8%
R2SNP7	10%	8%	8%
R2SNP8	96%	90%	92%
R2SNP9	12%	10%	10%
R2SNP10	36%	26%	30%
Others	5%	5%	5%
Interactions:			
IVY/ R2SNP5	100%	100%	100%
IVY/ R2SNP6	100%	99%	100%
IVY/ R2SNP7	100%	99%	100%
IVY/ R2SNP8	100%	89%	100%
IVY/ R2SNP9	100%	100%	100%
IVY/ R2SNP10	100%	96%	100%
Others	5%	5%	5%

Note: per-comparison type I error is controlled at 5%.