Published in final edited form as:
*Hum Hered*. 2014 ; 78(2): 104–116. doi:10.1159/000363660.

# A New System Identification Approach to Identifying Genetic Variants in Sequencing Studies for A Binary Phenotype

**Guolian Kang**[1,*,#], **Wenjian Bi**[2,*], **Yanlong Zhao**[2,#], **Ji-Feng Zhang**[2], **Jun J. Yang**[3], **Heng Xu**[3], **Mignon L. Loh**[4], **Stephen P. Hunger**[5], **Mary V. Relling**[3], **Stanley Pounds**[1], and **Cheng Cheng**[1]

[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

[2]Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

[4]Benioff Children's Hospital, University of California at San Francisco, San Francisco, California 94143, USA

[5]University of Colorado School of Medicine and Children's Hospital Colorado, Aurora, Colorado 80045, USA

## Abstract

We propose in this paper a set-valued (**SV**) system model, which is a generalized form of Logistic (**LG**) and Probit (**Probit**) regression, to be considered as a method for discovering genetic variants, especially rare genetic variants in next generation sequencing studies, for a binary phenotype. We propose a new set-valued system identification method to estimate all the underlying key system parameters for the **Probit** model and compare it with the **LG** model in the setting of genetic association studies. Across an extensive series of simulation studies, the **Probit** method maintained Type I error control and had similar or greater power than the **LG** method which is robust to different distributions of noise: logistic, normal or *t* distributions. Additionally, the **Probit** association parameter estimate was 2.7–46.8 fold less variable than the **LG** log-odds ratio association parameter estimate. Less variability in the association parameter estimate translates to greater power and robustness across the spectrum of minor allele frequencies (MAFs), and these advantages are the most pronounced for rare variants. For instance, in a simulation that generated data from an additive logistic model with odds ratio of 7.4 for a rare single nucleotide polymorphism with a MAF of 0.005 and a sample size of 2300, the **Probit** method had 60% power whereas the **LG** method had 25% power at the $\alpha=10^{-6}$ level. Consistent with these simulation results, the set of variants identified by the **LG** method was a subset of those

[#] Address for correspondence: Guolian Kang, Ph.D., Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA, Phone: +1-901-595-2666, Fax: +1-901-595-8843, Guolian.Kang@stjude.org or Yanlong Zhao, Ph.D., Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PRC, Phone: 86-10-62651446, Fax: 86-10-62587343, ylzhao@amss.ac.cn.

[*] These two authors contributed equally to this work.

identified by the **Probit** method in two example analyses. Thus, we suggest the **Probit** method may be a competitive alternative to the **LG** method in genetic association studies such as candidate gene, genome-wide, or next generation sequencing studies for a binary phenotype.

**Keywords**

Set-valued system model; binary phenotype; threshold model; genetic variants; rare variants; next-generation sequencing studies

## Introduction

For the past 10 years, genome-wide association studies (GWAS) have been an effective and successful method to detect common genetic variations associated with various phenotypes [1–3]. To date, the majority of phenotypes studied have been binary/categorical, continuous, or survival phenotypes. The standard linear regression model is the main method to analyze continuous phenotypes if the normality assumption holds approximately for its original or transformed scale. The Cox proportional hazard regression model is the key method to analyze the survival outcomes if the proportional hazard assumption holds approximately.

The logistic regression (**LG**) model is widely used to analyze the binary/categorical phenotype in GWAS. Often a binary phenotype is derived from a continuous variable by splitting the range at some threshold and categorizing individuals above and below that point into 2 separate groups of “affected” and “unaffected.” Examples of such designations include obesity defined based on body index mass [4], hypertension defined based on systolic blood pressure and/or diastolic blood pressure [5], and diabetes defined based on hemoglobin A1c level [6]. Moreover, some binary phenotypes may manifest from complicated unobserved or unobservable continuous variables such as expression of an unknown protein in a particular organ that causes the disease. Therefore, the simple **LG** model may be too naive to adequately reflect the underlying biology, resulting in performance reduction in studies of phenotypes as such.

The Cochran-Armitage trend test (CATT) is a widely used test for the binary phenotype (case-control) which assumes an additive mode of inheritance. CATT is equivalent to the score test for a logistic regression and has high power for additive and multiplicative disease models but much lower power for the recessive disease model [7–10]. The genotypic association test using Pearson’s Chi-square test for a 2×3 contingency table is robust for different disease models [11] but generally has lower power than CATT for the additive disease model. MAX3 is another widely used method that is robust for different disease models [12]. MAX3 is the maximum of the absolute values of CATT test statistics, assuming the additive, dominant and recessive disease modes of inheritance. The p-value of MAX3 can be estimated by the approximation method implemented in the software [13] because of its complicated asymptotic distribution. Other innovative methods include the entropy-based method which is generally as good as or more powerful than the genotypic association test [11, 14], and some genetic model selection and genetic model exclusion methods based on Hardy Weinberg disequilibrium information [15–17]. Although these methods have some advantages in certain situations, they either cannot adjust for confounding factors such as genetic ancestry, which is commonly adjusted in genetic association studies as the **LG** model does; or are time consuming when applied to GWAS; or can have lower power than CATT when the underlying disease model is additive or multiplicative [18], especially for a small sample size.

Set-valued system widely exists in reality. And in some cases, it can degenerate to the well-known threshold model. The corresponding set-valued system identification was first investigated for sensor systems [19]. In contrast to the traditional system identification method, set-valued system identification can estimate the model by set-valued information rather than the precise output. Although set-valued system identification is technically challenging, it has been successfully used in various fields such as sensor networks and telecommunications [20–21]. If the outcome is a linear function of covariates and the noise follows a normal distribution, then the set-valued system model is referred to as the **Probit** model, which is a viable choice for genetic association studies. If the noise follows a logistic distribution, then the set-valued system model is referred to as the logistic model, which is widely used in genetic association studies.

While it is widely believed that **LG** and **Probit** give very similar statistical analysis results in most applications because the cumulative distribution function of the standard normal distribution is very similar to that of the logistic distribution with mean 0 and scale 1 [22], some published research suggests that there may be some GWAS applications in which the two models have very different statistical properties. The **LG** model can be poorer than the Probit model in terms of goodness of fit in small sample size settings because the logistic distribution has heavier tails than the normal distribution [23]. Also, linear discriminate analysis, which like Probit regression is based on a normal distribution instead of the logistic distribution, has much greater asymptotic efficiency than **LG** regression [24]. Furthermore, it has been shown that a toxicology study that evaluates a binary response at three drug dose levels (−0.79, +0.79, and +2.69 on an arbitrary scale) with most subjects assigned to one of the extreme doses is optimal for differentiating between a **LG** and **Probit** regression model in terms of goodness of fit [25]. Furthermore, since changes of +/− 0.2 in the dose levels have minimal impact on the power to distinguish between the logistic and probit models [24] and the statistical results are invariant to shift and scale transformations of the dose levels, the result holds for transformed dose levels 0, 1, and 2, which is equivalent to the common representation of genotype data as the number of copies of the minor alleles. This result suggests that the logistic and probit models may have very different performance in terms of Type I error control and power in the analysis of the association of rare variants with disease status.

In this study, we propose a set-valued (**SV**) system model, which considers the dichotomization process of continuous phenotypes to model the relationship between the binary outcome and possible genetic or non-genetic explanatory factors in GWAS or next-generation sequencing (NGS) studies. We propose a set-valued system identification approach for the **Probit** model to estimate the parameter of interest and use a Wald test statistic for testing the null hypothesis of no genotype-binary phenotype association. We performed extensive simulation studies to compare the type I error rate and power of the

**Probit** and **LG** methods. Finally, we applied both methods to a mini-exome sequencing data set and a candidate gene study.

## Materials and methods

### Notation

We assume that there are $N_0$ cases and $N_1$ controls in a case-control genetic association study (total sample size $N = N_0+N_1$) and that the genetic polymorphism of interest is diallelic [e.g., single nucleotide polymorphism (SNP)]. The 2 alleles at a SNP are denoted as A and a, where A is the minor allele. The three genotypes are therefore AA, Aa, and aa. Suppose that observations ($s_i$, $X_i$, $G_i$) are available for $N$ individuals, $i = 1, 2, \ldots N$, where $\boldsymbol{s}_i$ is the case-control status of the subject $i$; $X_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}}$ is the vector of $m$ covariates that we need to adjust for (e.g., demographic or clinical variables); and $G_i = 0$, 1, or 2 is the numerical coding of the 3 genotype aa, Aa or AA of the SNP for the same individual.

### Logistic Regression (LG) Model

The **LG** model [26] commonly used to test association between SNP and a binary phenotype with adjustment for some covariates is

$$\mathrm{logit}\Pr(s_i{=}1){=}\alpha_0{+}\theta G_i{+}\gamma^T X_i, i{=}1,2,\ldots N \quad (1)$$

where $\alpha_0$ is an intercept term, $\theta$ is the regression coefficient for the SNP and $\gamma = [\gamma_1, \gamma_2, .., \gamma_{\mathrm{m}}]^T$ is a vector of regression coefficients for $m$ covariates. The above equation is equivalent to $\Pr(s{=}1){=}\frac{1}{1+\mathrm{e}^{-(\alpha_0+\theta G_i+\gamma^T X_i)}}$. Evaluating whether the genetic variant SNP influences the phenotype, adjusting for covariates, corresponds to testing the null hypothesis of $H_0$: $\theta = 0$. The Wald test will be used to test for the null hypothesis in order to be consistent with the **Probit** method below.

### The Set-Valued (SV) Model

Instead of directly modeling the relationship between the genetic variant and the phenotype using the logistic regression, we propose an new **SV** system model in which the observation of cases and controls are measured by a set-valued sensor [20–21]

$$\begin{cases} y{=}f(G,X){+}e, \\ s{=}I(y \in A), \end{cases}$$

where $I(y \in A)I$ is the indicate function of the set $A$, $f$ is a deterministic function of $G$ and $X$, $y$ is a latent continuous variable, and $e$ is the random noise. The most common simplified case of set-valued sensor is to introduce a threshold $c$ to dichotomize the continuous variable. In this case, the **SV** model is very similar to the well-known threshold model.

$$\begin{cases} y=f(G,X)+e, \\ s=I(y>c). \end{cases}$$

Furthermore, when the function *f* is a linear function of *G* and *X* and *e* follows a normal distribution, the **SV** model becomes the **Probit** model [26]

$$\begin{cases} y_i=\alpha_0+\theta G_i+\gamma^T X_i+e_i, \\ s_i=I(y_i>c), \quad i=1,2,\dots N, \end{cases} \quad (2)$$

where $y_i$ is a latent continuous variable that can be dichotomized as case/control, $e_i$ is the independent and identically distributed random noise which follows a normal distribution with a mean of 0 and a variance of $\sigma^2$, *c* is the threshold used to define case/control status, and the observation $s_i$ is determined by a threshold *c* and the latent variable $y_i$. The null hypothesis of $H_0$: $\theta = 0$ corresponds to no genetic effect of the SNP on the phenotype. The parameter $\theta$ is to be identified to test for the null hypothesis using the expectation-maximization (EM) algorithm below.

Note if the function *f* is a linear function of *G* and *X* but *e* follows a logistic distribution with a location 0 and a scale of 1, then the **SV** model becomes the **LG** model:

$$\Pr(s=1)=\frac{1}{1+\mathrm{e}^{\tilde{\alpha}+\tilde{\theta}G+\tilde{\gamma}^T X}}=\frac{1}{1+\mathrm{e}^{\mathrm{c}-(\alpha_0+\theta G_i+\gamma^T X_i)}},$$

If *c*=0, then it becomes the **LG** model (1).

Similarly, it can be seen that

$$\Pr(s=1)=\Phi(\tilde{\alpha}+\tilde{\theta}G+\tilde{\gamma}^T X),$$

where Φ is the cumulative distribution function for standard normal distribution, $\tilde{\alpha}=-\frac{c-\alpha_0}{\sigma}$, $\tilde{\theta}=\frac{\theta}{\sigma}$, $\tilde{\gamma}=\frac{\gamma}{\sigma}$. However, an important deviation from the usual probit regression modeling is that here we take a novel system identification approach to estimate all the key underlying system parameters $\theta$, $\gamma$, and *c* (see below). We call (2) the **Probit** model but coupled with the new algorithm in the remaining of the paper. The core algorithm of the system identification is the EM algorithm, instead of the traditional Newton-like method which is widely used in the usual probit regression method. EM algorithm has outstanding robustness and without calculating the Hessian-like matrix for every iteration step, the EM algorithm takes much less time per iteration. Hence, we expect that this approach will be more efficient when the binary observations approximately follow model (2). The details about the algorithm and the efficiency discussions can be seen in the Discussions, Supplementary Section 1 and Supplementary Table S2.

### Estimation of $\theta$ and the Test Statistic

The system parameters in (1) can be estimated by maximum likelihood through the EM algorithm [21]. Denote the vector of parameters ($\theta$, $\gamma^T$, $c$) by Θ, the vector ($G$, $X$,−1) by $\varphi$, and the maximum likelihood estimator of Θ by $\hat{\Theta}$. The iteration process of the EM-based system identification and the fisher Fisher information matrix of Θ at $\hat{\Theta}$, denoted as i($\hat{\Theta}$), can be obtained as (for details see Supplementary Section 1)

$$\hat{\Theta}^{k+1}=\hat{\Theta}^{k}+\left(\sum_{i=1}^{N}\varphi_i\cdot\varphi_i{}^{\mathrm{T}}\right)^{-1}\left[\sum_{i=1}^{N}f(\hat{a}_i^k)\left(\frac{I(s_i=1)}{1-F(\hat{a}_i^k)}-\frac{I(s_i=0)}{F(\hat{a}_i^k)}\right)\cdot\sigma^2\varphi_i\right]$$

and

$$\mathrm{i}(\hat{\Theta})=-E\left[\frac{\partial^2}{\partial\Theta^2}\mathrm{log}L(\Theta)|\hat{\Theta}\right]=\sum_{\mathrm{i}=1}^{N}\frac{1}{F(1-F)}f^2\cdot\varphi_i\cdot\varphi_i{}^{\mathrm{T}},$$

where $L$(Θ) is the likelihood function, F(.) and f(.) are the cumulative distribution function and probability distribution function of a normal distribution with mean 0 and variance $\sigma^2$, respectively. Testing for no genetic effect of SNP on the phenotype, that is, $H_0$: $\theta = 0$, can be constructed for the **Probit** method from the Wald statistic

$$W=\frac{\hat{\theta}^2}{\mathrm{i}(\hat{\Theta})^{-1}[1,1]},$$

where i($\hat{\Theta}$)$^{-1}$ [1,1] is the variance of $\hat{\theta}$, that is, the element at the first row and column of the inverse Fisher information matrix. Asymptotically, for large sample sizes, $W$ is distributed approximately as a central $\chi^2$ distribution with 1 degree of freedom under the null hypothesis of no association of SNP and phenotype.

## Simulations

### Data Generation

Simulation studies were performed to compare the relative performance of the **Probit** method coupled with the proposed EM algorithm and the **LG** method. In these simulations, given the minor allele frequency (MAF) $p_A$, the genotype frequencies $p(G=g)$ were calculated according to Hardy–Weinberg equilibrium (HWE) law, that is, $p(G=0)=(1–p_A)^2$, $p(G=1)=2p_A\ (1-p_A)$, $p(G=2)=(p_A)^2$. Two covariates were considered: $x_1$ was a binary variable that is 1 with a probability of 0.5 and 0 otherwise, and $x_2$ was a continuous variable that follows a standard normal distribution. The genotypes and 2 covariates for a population of 2,000,000 individuals were independently generated from their respective distributions.

The case-control status was determined from the generated genotype and covariate data according to two models respectively:

1. Logistic regression model (**LGsimu**):

   $\Pr(s_i=1|G_i,\ x_{i1},\ x_{i2})=\frac{\exp(\alpha_0+\theta G_i+0.5x_{i1}+0.5x_{i2})}{1+\exp(\alpha_0+\theta G_i+0.5x_{i1}+0.5x_{i2})}$, where $\alpha_0 = -2.2$ is the parameter chosen so that the disease prevalence is 0.1 among the subpopulation with $x_1 = x_2 = G = 0$.

2. Probit regression model (**PRsimu**): First a continuous variable was generated from $y_i = \alpha_0 + \theta G_i + 0.5x_{i1} + 0.5x_{i2} + e_i$, where $e_i$ follows the standard normal distribution and $\alpha_0 = -2.2$. Then, the individuals with a large value of $y_i$ greater than the threshold $c = \Phi^{-1}(1 - 0.1) + \alpha_0$ were declared as cases and the remaining individuals as controls. This model also gives a disease prevalence of 0.1 among the subpopulation with $x_1 = x_2 = G = 0$.

Then, $n$ cases and $n$ controls were randomly selected from a population of 2,000,000 individuals.

### Type I Error Rate Simulations

Eight values for MAFs of SNPs were considered: 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5. The case-control status was determined from the generated genotype and covariate data by using the two models mentioned above, with $\theta = 0$. To estimate the type I error rate of the **Probit** and **LG** method, 10,000,000 replicated datasets were simulated for each case-control study, with a small sample size of 500 cases and 500 controls and a large sample size of 2000 cases and 2000 controls for larger significant levels $\alpha = 0.05$ or 0.01 and stringent genome-wide levels $\alpha = 10^{-5}$ or $10^{-6}$ under the null hypothesis of $H_0$: $\theta = 0$, respectively.

### Power Simulations

Three genetic disease models were considered: additive, dominant, and recessive. The case-control status was determined from the generated genotype and covariate data according to the simulation methods given above, with $\theta$ varying from 0.3 to 2 at an increment of 0.1. Datasets were generated 10,000 times for each configuration. The **LG** and **Probit** methods used for the type I error simulations were applied to each data-set, and power was estimated as the proportion of $p$-values less than $\alpha = 10^{-6}$.

## Simulation Results

### Type I Error Rate

Table 1 shows empirical type I error rates estimated for both the **LG** and **Probit** methods. Regardless of significance levels, both methods correctly maintained type I error rates at the given levels but both are conservative if SNPs are rare and the sample size is small because of large variance of parameter estimate (Table 2 and Supplementary Figure S1).

### Power of the LG and Probit methods

Figures 1–2 show the power of the **LG** and **Probit** methods as a function of effect size ($\theta$) for additive and dominant disease models for $n$= 500 and 1000. As expected, the power of both methods increased with the increase in effect size. For a common SNP with an MAF of 0.2 or 0.05, the estimated coefficient of the Probit method was 1.6 to 1.8 times that of the LG method (Table 2), which is the case when both models fit well [22], so it is not surprising that the power of the **Probit** method was almost identical to or slightly greater than that of the **LG** method (Figures 1 and 2),, regardless of effect sizes of SNP ($\theta$) and the genetic disease model. For a rare SNP with an MAF of 0.01 or 0.005, the power of the **Probit** method was much greater than that of the **LG** method regardless of genetic disease models. The gain in efficiency for the new **Probit** method was noticeable in detecting rare variants with moderate sample sizes (Figures 1–3). If the phenotype was simulated using **LGsimu** under additive model, with a total sample size of 1,000, the power of the **Probit** method was 31% whereas that of the **LG** method was 0.05% for detecting a rare SNP with an MAF of 0.01 and an effect size of 2 (Figure 1B). For a total sample size of 2,000, the power of the **Probit** method was 34% whereas that of the **LG** method was only 6% for detecting a rare SNP with an MAF of 0.005 and an effect size of 2 (Figure 1D). If the phenotype was simulated using **PRsimu** under additive model, with a total sample size of 1,000, the power of the **Probit** method was 81% whereas that of the **LG** method was 43% for detecting a rare SNP with an MAF of 0.01 and an effect size of 1.8 (Figure 2B). For a total sample size of 2,000, the power of the **Probit** method was 77% whereas that of the **LG** method was only 39% for detecting a rare SNP with an MAF of 0.005 and an effect size of 1.8 (Figure 2D).

Figure 3 display the power of the **LG** and **Probit** methods as a function of sample size for the additive and dominant disease models. As expected, the power of both **LG** and **Probit** methods increased with an increase in sample size. For a common SNP with an MAF of 0.2 or 0.05 and an effect size of 0.4 or 0.8, respectively, the power of the **Probit** method was almost identical to that of the **LG** method regardless of the phenotype simulation model, sample size, and disease model. For a rare SNP with an MAF of 0.01 or 0.005 and an effect size of 1.6 or 2, respectively, the power of the **Probit** method was generally greater than that of the **LG** method regardless of the phenotype simulation model and disease model. The power difference became larger with moderate sample sizes. If the phenotype was simulated using **LGsimu** under additive model, with a total sample size of 2,100, the power of the **Probit** method was 80% whereas that of the **LG** method was 73% for detecting a rare SNP with an MAF of 0.01 and an effect size of 1.6. For a total sample size of 2,700, the power of the **Probit** method was 81% but that of **LG** method was only 56% for detecting a rare SNP with an MAF of 0.005 and an effect size of 2 (Figure 3A). If the phenotype was simulated using **PRsimu** under additive model, with a total sample size of 1,000, the power of the **Probit** method was 65% whereas that of the **LG** method was 30% for detecting a rare SNP with an MAF of 0.01 and an effect size of 1.6. For a total sample size of 2,000, the power of the **Probit** method was 83% whereas that of **LG** method was only 48% for detecting a rare SNP with an MAF of 0.005 and an effect size of 2 (Figure 3B).

The relationship between all parameter values and ratio of the power of the **Probit** method to that of the **LG** method was quantified by simulation studies. The relative power for a wide range of parameter setups ($\theta(\neq 0)$, $n$, $p_A$, and **PRsimu** or **LGsimu**) was first obtained and linear regression analysis was then performed using the log relative power as the outcome variable and the true parameter values as explanatory variables. The estimated mean log relative power of **Probit** to **LG** for testing $H_0$: $\theta = 0$ was $2.90 - 0.85\log_{10}(n) - 2.47p_A + 0.13\theta - 0.11\boldsymbol{I}(\textbf{PRsimu})$. This indicates that the magnitude of $\theta$, sample size $N$, and MAF of the SNP $p_A$ play dominant role in the relative power for testing $H_0$: $\theta = 0$.

Next we study the performance of the **Probit** method compared to entropy-based method [11] by simulations. Parameter setups were the same as those for Figures 1–2 except here we did not include any covariates and we used sample sizes of $n$=1000 and 2000. Figure 4 displays the power difference between **Probit** and entropy-based method for three disease models. The power of **Probit** was greater than that of entropy-based method for additive and dominant disease models if MAFs of SNP were 0.2, 0.05 and 0.01 (Figure 4A–4B). If the sample size was 2000, then **Probit** outperformed the entropy-based method for a SNP with MAF of 0.005. The maximum of the absolute values of the power differences between **Probit** and entropy-based method for additive and dominant models was 0.1. However, for recessive disease model, the entropy-based method was dominant over **Probit**. The power difference between **Probit** and entropy-based method could be as large as 0.80 (Figure 4C). For SNPs with MAF 0.01 and 0.005, there is no power difference between two methods for recessive disease model because the power of two methods was 0 due to small sample size (Figure 4C). These results are in fact consistent with the comparisons between **LG** and entropy-based method in the literature [11, 14].

## Variance of the Genetic Association Parameter Estimate

To show how the MAF of a SNP, sample size, genetic disease model, and different distribution of noise affect the parameter estimate of **Probit** and **LG**, respectively; and how they affect the ultimate power or type one error rate of both methods, respectively, we have conducted a series of simulations with a small sample size $n$=500, and large sample sizes of $n$=2000, 3000, 5000 and 10000 to investigate the genetic association parameter estimates and variance of genetic association parameter estimate.

Table 2 and Supplementary Figure S1 show the mean estimates of the genetic effect size association parameter, averaged estimated asymptotic variances, and empirical variances for the **LG** and **Probit** methods. Data were generated using the same parameter setup as given in Table 1 and Figures 1–2, except sample sizes. If $\theta = 0$, regardless of the phenotype simulation model, the mean estimates with both the **LG** and **Probit** methods were close to 0, but estimates with the **Probit** method were closer to 0 than those with the **LG** method (Table 2, Supplementary Figure S1B). If $\theta \neq 0$, the mean estimates with the **LG** (**Probit**) method appeared to be much closer to the true parameter values than those with **Probit** (**LG**) if the phenotype was generated from the **LG** model using **LGsimu** (**Probit** model using **PRsimu**) (Table 2, Supplementary Figure S1B). On average, the parameter estimates by the **LG** method were 2.11 times greater than those by the **Probit** method for a small sample size of 500, but decreased to 1.83, 1.80, 1.77, and 1.74 for large sample sizes of

2000, 3000, 5000 and 10000, respectively (Table 2, Supplementary Figure S1A). As expected, with the increase of sample size, the estimates become more robust and closer to their true parameter values regardless of the estimation method (Table 2, Supplementary Figure S1A and S1B). Not surprisingly, more common the SNP and/or larger the sample size, smaller was the bias of its estimate.

The averaged estimated asymptotic variance for the parameter estimate appeared to be close to its empirical counterpart for both the **LG** and **Probit** methods for the common SNP (Table 2 and Supplementary Figure S1C and S1D). Interestingly, for a small sample size, the averaged estimated asymptotic variance for the **LG** method was much larger than (mean: 26.9, range: 0.7~470.6) its empirical counterpart, especially for rare SNPs (mean: 90, range: 7.6~470.6), but not for the **Probit** method (mean: 1.2, range: 0.6~2.9), regardless of the phenotype simulation model (Table 2, Supplementary Figure S1C and S1D). The association parameter estimate for the **Probit** method was 2.66–46.8 fold less variable than that for the **LG** method. On average, for a SNP with MAF of 0.2, the empirical variance obtained by the **LG** method was 3.48 (Range: 2.67~38) times than that obtained by the **Probit** method, which is close to the ratio of the variance of the logistic distribution with scale of 1, i.e., $\pi^2/3$, to that of the standard normal distribution, i.e., 1. However, this value increased to 4.94 for a small sample size of 500 (Table 2, supplementary Figure S1F). Interestingly, for a SNP with MAF of 0.005, the empirical variance obtained by the **LG** method was 24.3 (Range: 2.67~195) times than that obtained by the **Probit** method which is about 7 times larger than that for a SNP with MAF of 0.2. This value increased to 35.1 for a small sample size of 500 (Table 2, supplementary Figure S1F). All these simulation results obviously demonstrate that **Probit** can give more robust and much less variable parameter estimate than **LG**, especially dominant for small sample sizes and rare variants, which translate to higher power of **Probit** than that of **LG**.

## Application to the Mini-Exome Data of Genetic Analysis Workshop 17

To evaluate the performance of the **Probit** method coupled with the new proposed algorithm, we analyzed data from the Genetic Analysis Workshop 17 (GAW17) which contained “mini-exome” sequence genotype data of 24,487 SNPs in 3,205 genomic regions of 697 unrelated individuals provided by the 1000 Genome Project [27]. Three quantitative phenotypes were simulated from the normal distribution. Two quantitative phenotypes and one latent disease liability were influenced by 39 SNPs in 9 genes, 72 SNPs in 13 genes, and 51 SNPs in 15 genes, respectively. The third quantitative phenotype was influenced only by the environments and not genetic variants. One qualitative phenotype denoted by $Q_4$ was simulated based on the three quantitative phenotypes and the latent liability and the top 30% of the distribution was declared affected. Furthermore, 200 replicate datasets were generated for each phenotype, using one fixed genotype data. First, quality control analysis was performed on the SNPs and SNPs with MAFs less than 0.00075 or HWE test p-values less than 0.00001 were excluded. The 1st, 10th, 100th and 200th qualitative traits were used as our outcomes and included age, gender, and smoking status as covariates in both the **LG** and **Probit** models.

At a significance level of 0.0001, no SNP was statistically significant for both the **LG** and **Probit** methods for 10th replicate data. Both methods identified the same causal SNP (C13S523) for the 1st and 100th replicate data. The **Probit** method identified two causal SNPs (C13S523 and C13S522) but the **LG** method only identified the causal SNPC13S522 for the 200th replicate data (Table 3). Neither method identified no-causal SNPs at the significance level of 0.0001. On average, the **Probit** method identified fewer number of non-causal SNPs but the same number of causal SNPs associated with the qualitative phenotype than the **LG** method at significance levels of 0.05, 0.01, and 0.001 (Supplementary Table S1).

## Application to *ARID5B* Gene in Acute Lymphoblastic Leukemia

Acute lymphoblastic leukemia (ALL) is the most common type of cancer in children and has different incidence rates in different racial/ethnic groups [28]. Genetic variants in *ARID5B* associated with risk of ALL have been reported recently [29–30]. We analyzed *ARID5B* genetic polymorphisms in childhood ALL in two populations of white and Hispanic children [31]. 978 white and 330 Hispanic children enrolled on Children's Oncology Group clinical trials [32] and 1046 white controls from the Genetic Association Information Network schizophrenia cohort [33–34] and 541 Hispanic control from HapMap II, the Human Variation Panel and Mexican participants in the Genetics of Asthma in Latino Americans study [35] were genotyped using Affymetrix SNP Array 6.

After quality control analysis, 49 SNPs within 10kb upstream or downstream of the gene were included for association testing of SNP with ALL susceptibility. Table 3 shows the SNPs with p-values less than 0.001 (0.05/49). We can see that in whites based on the **LG** method, 6 of the 49 SNPs, rs10821936, rs10821938, rs10994982, rs7087125, rs7896246, and rs7923074, had the p-values less than 0.001, and rs2893881 had a p-value of 0.001024, close to 0.001. Based on the **Probit** method, all these seven SNPs had p-values less than 0.001 (Table 3). In Hispanics, both methods identified the same set of SNPs associated with ALL susceptibility (Table 3). At more liberal significance levels of 0.05 and 0.01, both methods identified the same set of SNPs too (data not shown).

## Discussion

With the availability of data from whole-genome sequencing and whole-exome sequencing studies in which moderate sample sizes are used due to the high cost of sequencing technology [36–37] or the rare diseases in cancer genomics studies such as pediatric cancers of retinoblastoma and Ewing's sarcoma [38–39], there is an increasing demand for the development of powerful and robust association testing procedures for identifying genetic variations associated with a binary phenotype of interest. In this study, we propose a new **SV** system model, which is a generalized form of logistic and probit regression models, to model the relationship between a binary phenotype and genetic variants and a novel set-valued system identification approach for the **Probit** model to identifying the parameters association of interest. We compare it with the **LG** model. Simulations and real data applications show that the power gain of **Probit compared to LG** for binary phenotypes is robust to the distributions of noise: logistic or normal distribution, and various genetic disease models, and that **Probit** generally outperforms the commonly used **LG**. Furthermore, we also compared the elapsed time between our new algorithm and the built-in command *glmfit(x, y, 'binomial', 'link', 'probit')* in Matlab using simulations. We found that on average, the elapsed time of our new algorithm took 0.06 (range: 0.003~0.1022) seconds less time than glmfit for one SNP. For GWAS and NGS, we usually test for ~$10^6$ and ~$4\times10^7$ SNPs, respectively, and then it will save ~17 hours and ~677 hours compared to glmfit in matlab, respectively. But we did not find the difference of computing time between our new algorithm and *glm* function in R [40]. In addition, **Probit** has greater power than entropy-based method for additive and dominant but not recessive disease models. However, to the best of our knowledge, there is no program available in entropy-based method to include covariates in genetic association studies. In conclusion, we recommend the use of the **Probit** method coupled with our new algorithm instead of the **LG** method, regardless of the distribution of noise, sample size, and effect size of associations between variants and disease of interest, to identify genetic variants, especially rare variants, in genetic association studies.

When we estimate the parameters using system identification method, we suppose that the variance of noise is known as 1 because we are interested in testing genotype-phenotype associations not estimating the effect size of association. In real data analysis, the true variance of noise is usually unknown and also may not be equal to 1 which will definitely affect the power of the **LG** and **Probit** methods. By simulations with noise following a normal distribution of $N(\mu,\sigma^2)$, where $\sigma^2 = 3$ and $\sigma^2 = 1/3$, not surprisingly, as the true variance of the noise is bigger (smaller) than 1, the power of both methods will decrease (increase). However, as expected, the power of the **Probit** method is still identical to or greater than that of the **LG** method (data not shown). Thus, conclusions about the relative performance of the **Probit** and **LG** methods in this study are also robust to the *t* distribution of the underlying noise. In addition, if we are interested in estimating the association effect size of SNP on the phenotype, the parameter of variance of noise can also be estimated along with other parameters using expectation conditional maximization algorithm [21].

Besides SNP-based analysis, the **Probit** model coupled with the new algorithm can also be applied to any biologically meaningful mutants and mutant sets. It can be applied to a multiallelic locus, and the somatic status of structural variants such as copy number variants, copy-neutral regions of loss of heterogeneity, inversions and translocations. For next generation sequencing studies that involve rare variants, due to lack of power for single-locus approach, **Probit** method can be extended to a multiple-locus such as haplotype-based, gene/set-based, and pathway-based approach for detecting rare variants. Furthermore, the proposed **SV** model focuses on a binary phenotype with one threshold. However, in real data analyses, especially in the field of pharmacogenomics, the outcome could be multiple ordinal categories such as dosing of drugs, adverse events scored on scales using ordinal values (1–5) according to the Common Terminology Criteria for Adverse Events developed by the National Cancer Institute, or the effect of treatment on disease (e.g., tumour response in which the change of tumour size is categorized as a complete response, partial response, stable disease or progressive disease) [39]. The concept of a multiple-input-multiple-output linear system with quantized outputs [21] can be applied in these cases and hence can provide a comprehensive framework for a wide variety of genetic association studies. Similar to **LG** method, **Probit** method also has lower power than entropy-based method if the unknown underlying disease model is recessive. The power of **Probit** can be made robust against the underlying disease models by a computationally intensive approach: take as the test statistic the maximum of the absolute values of z test statistics assuming the additive, dominant and recessive disease models in **Probit** model. Then an empirical p-value can be obtained by a re-sampling method. In the current study, we have only investigated which method of the **LG** and **Probit** performs better in terms of association testing but in the future we will determine which method performs better in terms of model fitting and prediction.

We have implemented the **Probit** model coupled with the proposed EM algorithm in an R package and Matlab codes, which are available for free download from http://www.stjuderesearch.org/site/depts/biostats/software. The method can be easily applied to any genetic association studies no matter candidate gene, GWAS or NGS studies for a binary phenotype.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research is funded by the American Lebanese and Syrian Associated Charities (ALSAC), the National Natural Science Foundation of China under grants 11171333 and 61134013, the NIH grants U01 GM92666 and P30 CA21765, and the NCI grants U10 CA98543, U10 CA98413 and U24 CA114766 to the Children's Oncology Group. SPH is the Ergen Family Chair in Pediatric Cancer. The Genetic Analysis Workshop is supported by the NIH grant R01 GM031575. We thank Dr. Vani Shanker for editing the manuscript and Deqing Pei for providing the ALL data. We thank two anonymous reviewers for their insightful comments which have significantly improved the manuscript.

## References

1. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007; 445:881–885. [PubMed: 17293876]
2. The Wellcome Trust Case Control Consortium (WTCCC). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
3. Yang JJ, Cheng C, Devidas M, Cao X, Campana D, Yang W, Fan Y, Neale G, Cox N, Scheet P, et al. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. Blood. 2012; 120(20):4197–4204. [PubMed: 23007406]
4. Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM, Hypponen E, Holst C, Valcarcel B, Thiering E, et al. A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Genet. 2012; 44:526–531. [PubMed: 22484627]
5. Adeyemo A, Gerry N, Chen G, Herbert A, Doumatey A. A Genome-Wide Association Study of Hypertension and Blood Pressure in African. PLoS Genet. 2009; 5(7):e1000564. [PubMed: 19609347]
6. Florez JC. A genome-wide association study of treated A1C. Diabetes. 2010; 59(2):332–334. [PubMed: 20103712]
7. Armitage P. Tests for linear trends in proportions and frequencies. Biometrics. 1955; 11:375–386.

8. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet. 2010; 86(1):6–22. [PubMed: 20074509]

9. Cochran WG. Some methods for strengthening the common 2 tests. Biometrics. 1954; 10:417–451.

10. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics. 1997; 53:1253–1261. [PubMed: 9423247]

11. Ruiz-Marín M, Matilla-García M, Cordoba JAG, Susillo-González JL, Romo-Astorga A, González-Pérez A, Ruiz A, Gayán J. An entropy test for single-locus genetic association analysis. BMC Genetics. 2010; 11:19. [PubMed: 20331859]

12. Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002; 53:146–152. (erratum in Hum Hered 2009; 68: 220). [PubMed: 12145550]

13. Zang Y, Fung WK, Zheng G. Simple algorithms to calculate asymptotic null distributions for robust tests in case-control genetic association studies in R. J Stat Softw. 2010; 33(8)

14. Kang G, Zuo Y. Entropy-based joint analysis for two-stage genomewide association studies. J Hum Genet. 2007; 52:747–756. [PubMed: 17687620]

15. Chen J, Chatterjee N. Exploiting hardy-weinberg equilibrium for efficient screening of single SNP associations from case-control studies. Human Hered. 2007; 63:196–204.

16. Joo J, Kwak M, Ahn K, Zheng G. A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium. Biometrics. 2009; 65:1115–1122. [PubMed: 19432787]

17. Zheng G, Ng HKT. Genetic model selection in two-phase analysis for case-control association studies. Biostatistics. 2008; 9:391–399. [PubMed: 18003629]

18. Joo J, Kwak M, Zheng G. Improving power for testing genetic association in case-control studies by reducing alternative space. Biometrics. 2009; 66:266–276. [PubMed: 19397584]

19. Wang L, Zhang J, Yin G. System identification using binary sensors. IEEE TAC. 2003; 48(11): 1892–1907.

20. Wang L, Yin G, Zhang J, Zhao Y. System identification with quantized observations. Birkhauser. 2010

21. Godoy B, Goodwin G, Aguero J, Marelli D, Wigren T. On identification of FIR systems having quantized output data. Automatica. 2011; 47:1905–1915.

22. Agresti, A. Categorical data analysis. Hoboken, NJ: John Wiley and Sons; 2002.

23. Cakmakyapan S, Goktas A. A comparison of binary logit and probit models with a simulation study. J Soc and Econ Stat. 2013; 2(1):1–17.

24. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. 1975; 70:892–898.

25. Chambers EA, Cox DR. Discrimination between alternative binary response models. Biometrika. 1967; 3(4):573–578. [PubMed: 6064019]

26. McCullagh, P.; Nelder, J. Generalized Linear Models. London: Chapman and Hall; 1989.

27. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

28. Pui CH, Evans WE. Treatment of acute lymphoblastic leukemia. N Engl J Med. 2006; 354:166–178. [PubMed: 16407512]

29. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JA, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet. 2009; 41:1006–1010. [PubMed: 19684604]

30. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, Willman C, Neale G, Downing J, Raimondi SC, Pui CH, Evans WE, Relling MV. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet. 2009; 41(9):1001–1005. [PubMed: 19684603]

31. Xu H, Cheng C, Devidas M, Pei D, Fan Y, Yang W, Neale G, Scheet P, Burchard EG, Torgerson DG, et al. ARID5B genetic polymorphisms contribute to racial disparities in the incidence and

treatment outcome of childhood acute lymphoblastic leukemia. J Clin Oncol. 2012; 30:751–757. [PubMed: 22291082]

32. Borowitz MJ, Devidas M, Hunger SP, Bowman WP, Carroll AJ, Carroll WL, Linda S, Martin PL, Pullen DJ, Viswanatha D, et al. Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: A Children's Oncology Group study. Blood. 2008; 111:5477–5485. [PubMed: 18388178]
33. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009; 460:753–757. [PubMed: 19571809]
34. Purcell SM, Wray NR, et al. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]
35. Burchard EG, Avila PC, Nazario S, Casal J, Torres A, Rodriguez-Santana JR, Toscano M, Sylvia JS, Alioto M, Salazar M, et al. Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. Am J Respir Crit Care Med. 2004; 169:386–392. [PubMed: 14617512]
36. Lanktree MB, Hegele RA, Schork NJ, Spence JD. Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. Circ Cardiovasc Genet. 2010; 3:215–221. [PubMed: 20407100]
37. Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, Wright FA, Rieder MJ, Tabor HK, Nickerson DA, et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet. 2012; 44(8): 886–889. [PubMed: 22772370]
38. Gurney JG, Severson RK, Davis S, Robison LL. Incidence of cancer in children in the United States. Sex-, race-, and 1-year age-specific rates by histologic type. Cancer. 1995; 75:2186–95. [PubMed: 7697611]
39. Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. Nat Rev Genet. 2013; 14(1):23–34. [PubMed: 23183705]

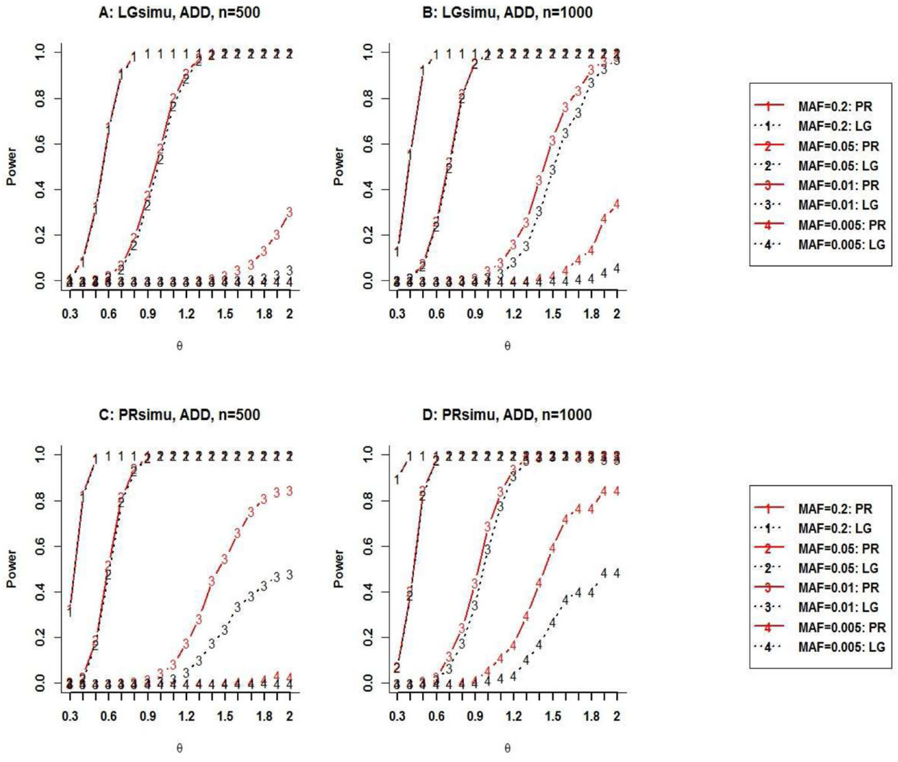

**Figure 1. Power of Probit and LG methods for the additive model using LGsimu and PRsimu**

The upper and lower panels showed results using **LGsimu** and **PRsimu**, respectively. A, C and B, D panels were for sample sizes of *n*=500 and *n*=1000, respectively. The solid and dotted lines corresponded to the **Probit (PR)** and **LG** methods, respectively. The numbers of 1–4 corresponded to MAFs of SNPs, 0.2, 0.05, 0.01 and 0.005, respectively. The significance level of the test was $1\times10^{-6}$.

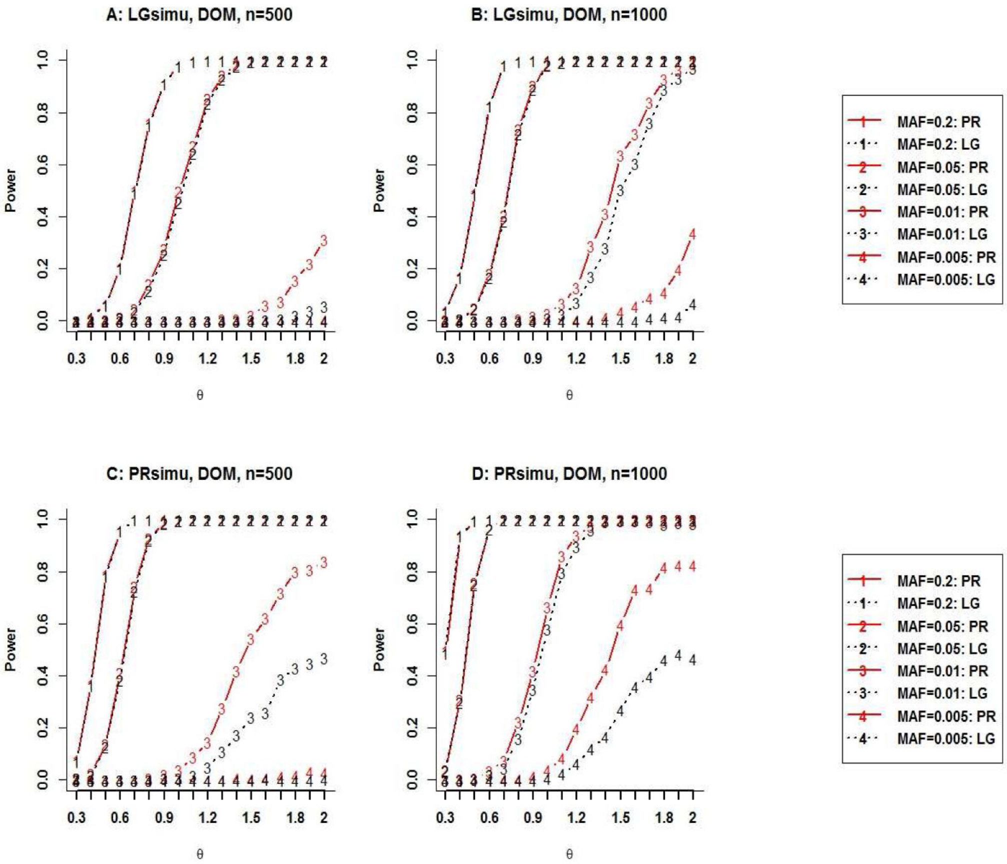

**Figure 2. Power of Probit and LG methods for the dominant model using LGsimu and PRsimu**

The upper and lower panels showed results using **LGsimu** and **PRsimu**, respectively. A, C and B, D panels were for sample sizes of *n*=500 and *n*=1000, respectively. The solid and dotted lines corresponded to the **Probit (PR)** and **LG** methods, respectively. The numbers of 1–4 corresponded to MAFs of SNPs, 0.2, 0.05, 0.01 and 0.005, respectively. The significance level of the test was $1\times10^{-6}$.

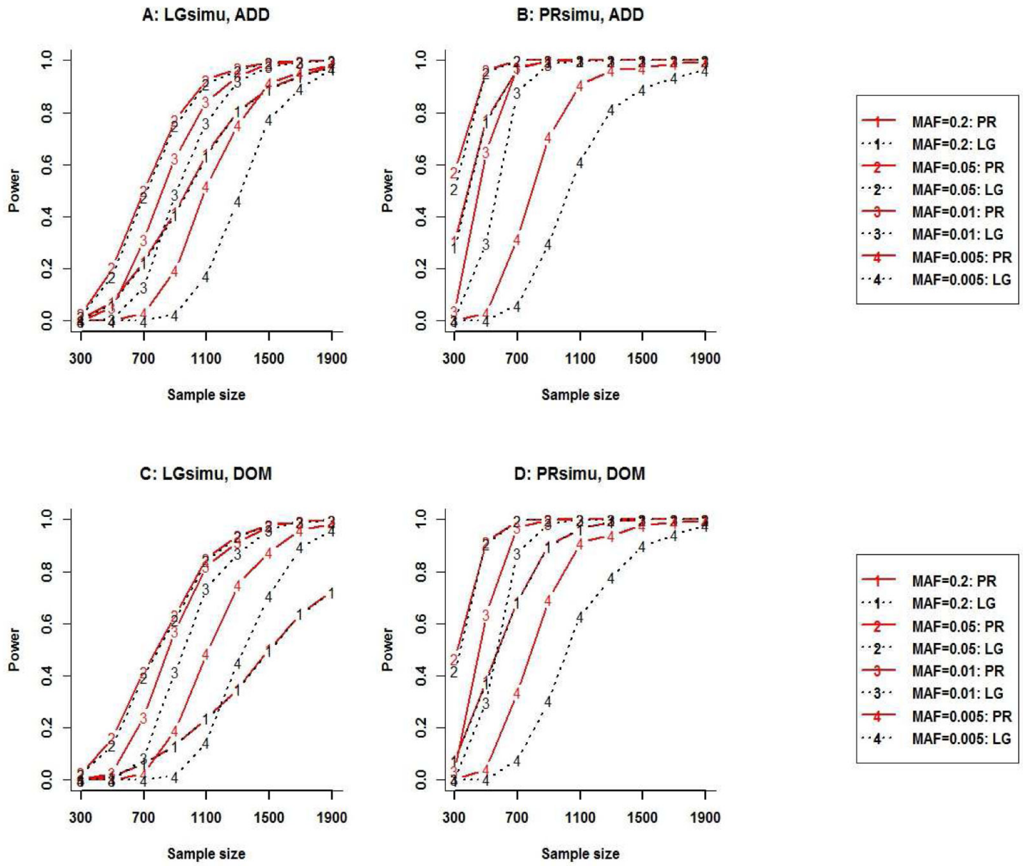

**Figure 3. Power of Probit and LG methods for the additive and dominant models as a function of sample size**

The upper and lower panels showed results for additive and dominant modelsPRsimu, respectively. The A, C and B, D panels corresponded to **LGsimu** and **PRsimu**, respectively. The solid and dotted lines correspond to the **Probit (PR)** and **LG** methods, respectively. The significance level of the test was $1\times10^{-6}$. $\beta_g$ values were 0.4, 0.8, 1.6 and 2 for SNPs with MAFs of 0.2, 0.05, 0.01 and 0.005, respectively.

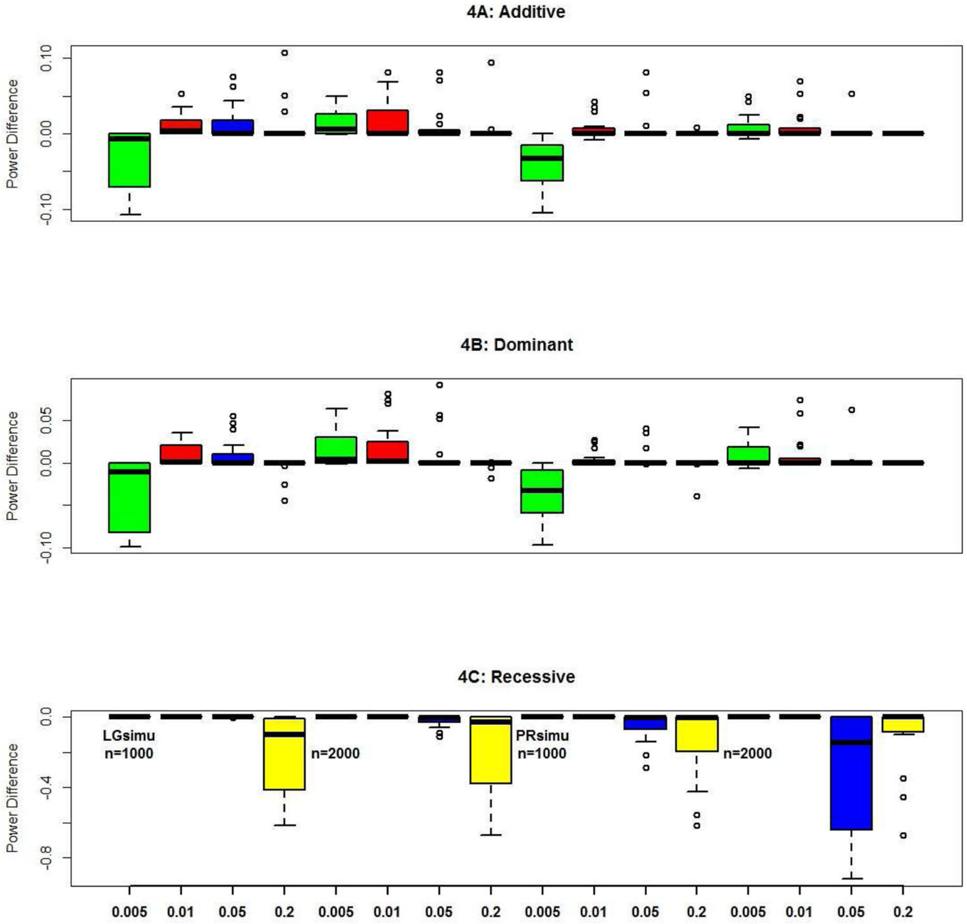

**Figure 4. Power difference between Probit and entropy-based method**

The A, B and C panels corresponded to additive, dominant and recessive disease models. All parameter setups were the same as Figures 1–2.

## Table 1

Type I error of the Probit and logistic regression (LG) methods

| | | | LG | | | | Probit | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***n*** | ***p*_A** | **SimuModel** | **0.05** | **0.01** | **$10^{-5}$** | **$10^{-6}$** | **0.05** | **0.01** | **$10^{-5}$** | **$10^{-6}$** |
| 500 | 0.005 | LGsimu | 0.0201 | 0.001 | 0 | 0 | 0.0332 | 0.0021 | 0 | 0 |
| 500 | 0.005 | PRsimu | 0.0453 | 0.0074 | 0 | 0 | 0.0488 | 0.0091 | 0 | 0 |
| 500 | 0.01 | LGsimu | 0.0257 | 0.004 | 0 | 0 | 0.0349 | 0.004 | $1.00\times10^{-6}$ | $1.00\times10^{-6}$ |
| 500 | 0.01 | PRsimu | 0.0459 | 0.0078 | 0 | 0 | 0.0484 | 0.0089 | $1.00\times10^{-6}$ | 0 |
| 500 | 0.05 | LGsimu | 0.041 | 0.0042 | 0 | 0 | 0.0475 | 0.0073 | 0 | 0 |
| 500 | 0.05 | PRsimu | 0.048 | 0.0087 | $2.00\times10^{-6}$ | 0 | 0.0493 | 0.0094 | $4.00\times10^{-6}$ | 0 |
| 500 | 0.2 | LGsimu | 0.0431 | 0.0053 | 0 | 0 | 0.0478 | 0.0074 | 0 | 0 |
| 500 | 0.2 | PRsimu | 0.0486 | 0.0092 | $4.00\times10^{-6}$ | 0 | 0.0494 | 0.0097 | $7.00\times10^{-6}$ | 0 |
| 2000 | 0.005 | LGsimu | 0.0487 | 0.009 | $5.00\times10^{-6}$ | 0 | 0.05 | 0.0096 | $6.00\times10^{-6}$ | $1.00\times10^{-6}$ |
| 2000 | 0.005 | PRsimu | 0.05 | 0.0096 | $6.00\times10^{-6}$ | $1.00\times10^{-6}$ | 0.0503 | 0.0098 | $7.00\times10^{-6}$ | $1.00\times10^{-6}$ |
| 2000 | 0.01 | LGsimu | 0.0489 | 0.009 | 0 | 0 | 0.0499 | 0.0096 | $2.00\times10^{-6}$ | 0 |
| 2000 | 0.01 | PRsimu | 0.0489 | 0.0096 | $7.00\times10^{-6}$ | 0 | 0.049 | 0.0097 | $8.00\times10^{-6}$ | 0 |
| 2000 | 0.05 | LGsimu | 0.05 | 0.0097 | $9.00\times10^{-6}$ | 0 | 0.0504 | 0.0099 | $1.10\times10^{-5}$ | 0 |
| 2000 | 0.05 | PRsimu | 0.05 | 0.0099 | $8.00\times10^{-6}$ | 0 | 0.05 | 0.0099 | $1.00\times10^{-5}$ | 0 |
| 2000 | 0.2 | LGsimu | 0.0501 | 0.0098 | $1.30\times10^{-5}$ | 0 | 0.0503 | 0.01 | $1.30\times10^{-5}$ | 0 |
| 2000 | 0.2 | PRsimu | 0.0499 | 0.01 | $9.00\times10^{-6}$ | 0 | 0.05 | 0.01 | $9.00\times10^{-6}$ | 0 |

## Table 2

Comparison of performances of the Probit and LG methods

| | | | | | LG | | | Probit | | | Ratio (LG/Probit) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $p_A$ | $\theta$ | Simulation Model | Disease Model | $\overline{\hat{\theta}}$[a] | $\overline{\hat{var}(\hat{\theta})}$[b] | var $(\hat{\theta})$[c] | $\overline{\hat{\theta}}$[a] | $\overline{\hat{var}(\hat{\theta})}$[b] | var $(\hat{\theta})$[c] | $\overline{\hat{\theta}}$[a] | var $(\hat{\theta})$[c] |
| 500 | 0.005 | 0 | LGsimu | H0 | −0.1296 | 67.125 | 4.3431 | −0.0456 | 0.8542 | 0.3906 | 2.84 | 11.12 |
| 500 | 0.005 | 0.5 | LGsimu | ADD | 0.6552 | 24.02 | 2.809 | 0.3473 | 0.4247 | 0.2837 | 1.89 | 9.903 |
| 500 | 0.005 | 2 | LGsimu | ADD | 2.88 | 589.52 | 11.216 | 1.334 | 0.7324 | 0.3975 | 2.16 | 28.21 |
| 500 | 0.005 | 0 | PRsimu | H0 | 0.01708 | 71.031 | 4.0925 | −0.0017 | 1.128 | 0.4007 | −10 | 10.21 |
| 500 | 0.005 | 0.5 | PRsimu | ADD | 1.077 | 29.91 | 2.7258 | 0.5824 | 0.415 | 0.2645 | 1.85 | 10.31 |
| 500 | 0.005 | 2 | PRsimu | ADD | 7.263 | 18442 | 39.188 | 2.407 | 2.506 | 0.8591 | 3.02 | 45.61 |
| 500 | 0.2 | 0 | LGsimu | H0 | 0.00732 | 0.0135 | 0.0147 | 0.0046 | 0.0052 | 0.0056 | 1.58 | 2.63 |
| 500 | 0.2 | 0.5 | LGsimu | ADD | 0.5079 | 0.013 | 0.013 | 0.3123 | 0.0048 | 0.0049 | 1.63 | 2.673 |
| 500 | 0.2 | 2 | LGsimu | ADD | 2.014 | 0.0207 | 0.0211 | 1.206 | 0.0067 | 0.007 | 1.67 | 2.998 |
| 500 | 0.2 | 0 | PRsimu | H0 | −0.005 | 0.016 | 0.017 | −0.0026 | 0.0058 | 0.0061 | 1.95 | 2.779 |
| 500 | 0.2 | 0.5 | PRsimu | ADD | 0.8689 | 0.0157 | 0.0153 | 0.5219 | 0.0055 | 0.0053 | 1.66 | 2.88 |
| 500 | 0.2 | 2 | PRsimu | ADD | 3.454 | 0.0443 | 0.0437 | 2.011 | 0.0125 | 0.0128 | 1.72 | 3.409 |
| 2000 | 0.005 | 0 | LGsimu | H0 | −0.0431 | 0.1151 | 0.1248 | −0.0265 | 0.0438 | 0.0471 | 1.63 | 2.649 |
| 2000 | 0.005 | 0.5 | LGsimu | ADD | 0.5008 | 0.0993 | 0.1023 | 0.3071 | 0.0371 | 0.038 | 1.63 | 2.691 |
| 2000 | 0.005 | 2 | LGsimu | ADD | 2.033 | 0.1134 | 0.1128 | 1.199 | 0.0333 | 0.0329 | 1.7 | 3.427 |
| 2000 | 0.005 | 0 | PRsimu | H0 | −0.0177 | 0.1351 | 0.1515 | −0.0108 | 0.0488 | 0.0541 | 1.64 | 2.797 |
| 2000 | 0.005 | 0.5 | PRsimu | ADD | 0.916 | 0.1113 | 0.1161 | 0.5485 | 0.0392 | 0.041 | 1.67 | 2.834 |
| 2000 | 0.005 | 2 | PRsimu | ADD | 3.649 | 2.1054 | 1.002 | 2.034 | 0.0748 | 0.0832 | 1.79 | 12.05 |
| 2000 | 0.2 | 0 | LGsimu | H0 | 0.00376 | 0.0034 | 0.003 | 0.0024 | 0.0013 | 0.0011 | 1.57 | 2.621 |
| 2000 | 0.2 | 0.5 | LGsimu | ADD | 0.503 | 0.0032 | 0.0033 | 0.3094 | 0.0012 | 0.0012 | 1.63 | 2.675 |
| 2000 | 0.2 | 2 | LGsimu | ADD | 2.004 | 0.0051 | 0.0053 | 1.201 | 0.0017 | 0.0018 | 1.67 | 2.96 |
| 2000 | 0.2 | 0 | PRsimu | H0 | −0.0091 | 0.004 | 0.004 | −0.0052 | 0.0014 | 0.0014 | 1.76 | 2.767 |
| 2000 | 0.2 | 0.5 | PRsimu | ADD | 0.8683 | 0.0039 | 0.0039 | 0.5218 | 0.0014 | 0.0014 | 1.66 | 2.832 |
| 2000 | 0.2 | 2 | PRsimu | ADD | 3.425 | 0.0109 | 0.0105 | 1.997 | 0.0031 | 0.0031 | 1.72 | 3.395 |

[a] The averaged estimate for 1000 replicates.

[b] The averaged estimated asymptotic variance for 1000 replicates.

[c] The empirical variance for 1000 replicates.

## Table 3

SNPs associated with $Q_4$ for GAW17 data with p-values less than $1 \times 10^{-4}$

| Phenotype dataset | SNP | MAF | Probit | LG |
|---|---|---|---|---|
| 1st | C13S523 | 0.067 | $6.68 \times 10^{-6}$ | $6.57 \times 10^{-6}$ |
| 100th | C13S523 | 0.067 | $2.10 \times 10^{-6}$ | $1.98 \times 10^{-6}$ |
| 200th | C13S522 | 0.03 | $6.16 \times 10^{-5}$ | $6.52 \times 10^{-5}$ |
|  | C13S523 | 0.067 | $7.13 \times 10^{-5}$ | 0.00011 |

MAF: minor allele frequency.

## Table 4

SNPs associated with ALL susceptibility in White and Hispanic

| | White | | | | Hispanic | | | |
|---|---|---|---|---|---|---|---|---|
| **SNP** | **MA** | **MAF** | **LG** | **Probit** | **MA** | **MAF** | **LG** | **Probit** |
| rs10821936 | C | 0.40 | $8.34\times10^{-20}$ | $2.70\times10^{-20}$ | T | 0.43 | $1.03\times10^{-7}$ | $6.99\times10^{-8}$ |
| rs10821938 | A | 0.48 | $1.47\times10^{-14}$ | $8.89\times10^{-15}$ | C | 0.37 | $4.27\times10^{-7}$ | $2.74\times10^{-7}$ |
| rs10994982 | G | 0.47 | $2.66\times10^{-7}$ | $2.41\times10^{-7}$ | G | 0.34 | $3.81\times10^{-6}$ | $3.47\times10^{-6}$ |
| rs7087125 | T | 0.5 | $9.22\times10^{-6}$ | $8.76\times10^{-6}$ | T | 0.45 | 0.75 | 0.75 |
| rs7896246 | A | 0.40 | $1.03\times10^{-19}$ | $3.32\times10^{-20}$ | G | 0.45 | $2.77\times10^{-7}$ | $2.24\times10^{-7}$ |
| rs7923074 | A | 0.48 | $1.50\times10^{-13}$ | $9.85\times10^{-14}$ | C | 0.37 | $2.09\times10^{-7}$ | $1.35\times10^{-7}$ |
| rs2893881 | G | 0.16 | 0.001 | 0.000996 | G | 0.32 | 0.007 | 0.006 |

MA: minor allele; MAF: minor allele frequency.