Published in final edited form as:
Genet Epidemiol. 2013 November ; 37(7): 686–694. doi:10.1002/gepi.21761.

A whole-genome simulator capable of modeling high-order epistasis for complex disease

Wei Yang[1] and **Charles Gu**[1,2]

[1]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO

[2]Department of Genetics, Washington University School of Medicine, St. Louis, MO

Abstract

Genome-wide association studies (GWAS) have been successful in finding numerous new risk variants for complex diseases, but the results almost exclusively rely on single-marker scans. Methods that can analyze joint effects of many variants in GWAS data are still being developed and trialed. To evaluate the performance of such methods it is essential to have a GWAS data simulator that can rapidly simulate a large number of samples, and capture key features of real GWAS data such as linkage disequilibrium (LD) among single-nucleotide polymorphisms (SNPs) and joint effects of multiple loci (multilocus epistasis). In the current study, we combine techniques for specifying high-order epistasis among risk SNPs with an existing program GWAsimulator[Li and Li 2008] to achieve rapid whole-genome simulation with accurate modeling of complex interactions. We considered various approaches to specifying interaction models including: departure from product of marginal effects for pair-wise interactions, product terms in logistic regression models for low-order interactions, and penetrance tables conforming to marginal effect constraints for high-order interactions or prescribing known biological interactions. Methods for conversion among different model specifications are developed using penetrance table as the fundamental characterization of disease models. The new program, called simGWA, is capable to efficiently generate large samples of GWAS data with high precision. We show that data simulated by simGWA are faithful to template LD structures, and conform to pre-specified diseases models with (or without) interactions.

Keywords

genome-wide simulation; epistasis; gene-gene interaction; genome-wide association

Introduction

In the past few years, waves of genome-wide association studies (GWAS) have identified numerous genetic risk factors of complex diseases[Hindorff, et al.; Manolio 2013; O'Seaghdha and Fox 2011; Willer and Mohlke 2012]. Although published GWAS analyses rely almost exclusively on single-marker scans, the importance of joint effects of multiple

Address Correspondence to: Dr. C. Charles Gu, Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 S. Euclid Avenue, St. Louis, MO 63110, Phone: (314) 362-3642, Fax: (314) 362-2693, gc@wubios.wustl.edu.

The authors declare that there is no conflict of interests.

risk variants is increasingly recognized by both methodologists and practitioners of the field[Aschard, et al. 2012; Cordell 2009; Kirino, et al. 2013; Lucas, et al. 2012; Ma, et al. 2012; MacLellan, et al. 2012; Pandey, et al. 2012]. Many statistical methods have been attempted for analyzing such joint effects, from multiple regression models, to haplotype-based tests, and to tests for interactions of multiple loci (within a genomic region or across the whole genome)[Gao, et al. 2013; Gyenesei, et al. 2012; Hahn, et al. 2003; Jin, et al. 2010; Oh, et al. 2012; Wan, et al. 2010; Wu, et al. 2010; Yang, et al. 2012; Yang, et al. 2011; Yang and Gu 2013]. To facilitate development of new statistical methods and better understand high-order gene-gene interactions (epistasis), it is essential to have a GWAS data simulator that not only can simulate huge amounts of genotype data with realistic genome-wide LD structure in reasonable computational time, but also can correctly model complex interactions among many risk loci.

A rapid whole-genome simulator called GWAsimulator[Li and Li 2008] does a great job in simulating marginal effects using “retrospective sampling” [Durrant, et al. 2004] (first sampling genotypes at risk loci conditional on disease status, then generating haplotypes by a moving-window algorithm). The algorithm simply copies and binds small pieces of haplotypes templates from real-world populations, making it possible to generate very large data sets in reasonable time, and at the same time, making it capable of simulating realistic LD structure by using real-population haplotype templates such as those from the HapMap project[Gibbs, et al. 2003]. However, the algorithm lacks a way to correctly handle multilocus interactions, which restricts its utility in studying complex diseases models with complex interactions.

A general way to specify higher order effects of multiple SNPs is to use a penetrance table. The penetrance at the risk locus is the conditional probability that an individual with a given genotype is affected by the disease of interest. A penetrance table consists of penetrance values for all possible combinations of multilocus risk genotypes, which, in conjunction with the genotype frequencies, fully characterizes the joint distribution of disease status and genotypes at the risk loci. We had previously developed a novel method[Yang and Gu 2008] for generating complex penetrance tables involving high order interactions for any given set of marginal effects of risk loci and risk alleles frequencies.

In the present study, we combine improved interaction model specification with GWAsimulator’s retrospective sampling to achieve rapid whole-genome simulation of GWAS data with accurate modeling of complex interactions. The resulting new algorithm is implemented in an R package called simGWA, which allows convenient and accurate specification of high-order effects using various approaches, including (A) departure from product of marginal effects for pair-wise interactions, (B) logistic regression models for low-order interactions, (C) penetrance tables conforming to given marginal constraints for high-order interactions, and (D) special penetrance tables prescribing known biological interactions. Penetrance table is used as the canonical characterization of complex disease models and to simulate genotypes at the disease loci. We will first introduce how penetrance tables are used in the simulation. Then, we describe GWAsimulator’s disease modeling and its limitation in specifying interactions, followed by description of our approaches to

generate correct penetrance tables for complex disease models. Finally we give an overview of the implementation of simGWA and the evaluation of its performance.

Methods

Flexible choice of disease models can be achieved by using penetrance tables to specify how combinations of risk SNPs affect the disease status.

Assume that the disease prevalence is K, and it involves m disease loci. At locus i (i = 1,2, …,m), g_i (=0, 1 or 2) is the risk allele count and p_i is the risk allele frequency. For a multilocus genotype combination $G = (g_1,g_2,\ldots,g_m)$, denote the penetrance by $f(G)$ = Pr(affected | G), which is the probability of being affected conditional on genotype G. Then, for a case subject, the probability that it has genotype G_0 is

$$\Pr(G_0|\text{affected})=\frac{\Pr(\text{affected}|G_0)\times\Pr(G_0)}{\sum_{G=\{g_1,\ldots,g_m\}}\Pr(\text{affected}|G)\times\Pr(G)}=\frac{f(G_0)\Pr(G_0)}{\sum_{G=\{g_1,\ldots,g_m\}}f(G)\times\Pr(G)} \quad (1)$$

For a control subject,

$$\Pr(G_0|\text{unaffected})=\frac{\Pr(\text{unaffected}|G_0)\times\Pr(G_0)}{\sum_{G=\{g_1,\ldots,g_m\}}\Pr(\text{unaffected}|G)\times\Pr(G)}=\frac{(1-f(G_0))\Pr(G_0)}{\sum_{G=\{g_1,\ldots,g_m\}}(1-f(G))\Pr(G)} \quad (2)$$

The denominators in the above formulae are summed over all possible genotypes comprising the m disease loci. Under assumptions of Hardy-Weinberg equilibrium and that all disease loci are unlinked, the probability of a genotype G is calculated as

$$\Pr(G)=\prod_{i=1}^{m}\Pr(g_i)=\prod_{i=1}^{m}[1+I(g_i=1)]p_i^{g_i}(1-p_i)^{2-g_i} \quad (3)$$

Thus, given the full penetrance table $f(\cdot)$ and allele frequencies p_i of all risk loci, it is straightforward to determine the distribution of genotypes in cases and controls using equations (1) and (2).

These equations form the basis for “retrospective sampling” (sampling genotypes or haplotypes of subjects conditional on the disease status) used by GWAsimulator[Li and Li 2008]. It is more efficient than a “prospective” approach, where multisite joint disease genotypes are randomly generated, but only accepted at a probability equal to the penetrance of that genotype combination[Peng and Amos 2010; Pinelli, et al. 2012].

After determining genotypes at the disease loci from retrospective sampling, genotypes at neighboring positions on the same chromosome are simulated one-by-one using a moving-window algorithm[Durrant, et al. 2004]. Simply speaking, for each partially simulated haplotype, the algorithm first finds haplotypes that match the already simulated haplotype in a small window among the haplotype templates. In these matched template haplotypes, the alleles at the next un-simulated position are counted and used to get the probability to

simulate the allele at this position. This process is then repeated again to get the next allele on the template, until the whole chromosome is simulated.

Using templates ensures that the simulated chromosomes bear LD structures similar to the population of interest. Such templates may be generated from existing GWAS data, or easily retrieved from the HapMap[Gibbs, et al. 2003] website, where phased data are provided for Caucasians (CEU data set), Africans (YRI data set) and East Asians (CHB+JPT). Naturally, the accuracy and resolution of the simulated LD structures depend on the sample sizes and data quality of the original template-generating samples.

Model specification used by GWAsimulator

GWAsimulator assumes that the penetrances link to the genotypes through a logistic function. Specifically, assuming there are *m* risk SNPs, when there are no interactions, the logit function of penetrances is

$$\text{logit}(f(G))=\text{logit}(\Pr(\text{affected}|g_1,\dots g_m))=\alpha+\beta_{11}I_{\{g_1=1\}}+\beta_{12}I_{\{g_1=2\}}+\cdots+\beta_{m1}I_{\{g_m=1\}}+\beta_{m2}I_{\{g_m=2\}} \quad (4)$$

and when pair-wise interactions exist between some SNP pairs, it is

$$\begin{aligned}\text{logit}(f(G))&=\text{logit}(\Pr(\text{affected}|g_1,\dots g_m))\\&=\alpha+\sum_{i=1}^{m}[\beta_{i1}I_{\{g_i=1\}}\\&\quad+\beta_{i2}I_{\{g_i=2\}}]+\sum_{k,l}[\gamma_{kl,11}I_{\{g_k=1,g_l=1\}}\\&\quad+\gamma_{kl,12}I_{\{g_k=1,g_l=2\}}\\&\quad+\gamma_{kl,21}I_{\{g_k=2,g_l=1\}}\\&\quad+\gamma_{kl,22}I_{\{g_k=2,g_l=2\}}]\end{aligned} \quad (5)$$

where logit(x) = x / (1−x); α is the constant coefficient; β_{i1} and β_{i2} are the coefficients for the effects of having 1 copy of targeted allele and 2 copies of the allele at the *i* th risk SNP, respectively; g_i is the number of copies of the targeted allele at SNP *i*, and $I_{\{g_i=n\}}$ is an indicator function of whether the copies of the allele is *n*; $\gamma_{kl,11}$, $\gamma_{kl,12}$, $\gamma_{kl,21}$, and $\gamma_{kl,21}$ are coefficients for the interaction between SNP *k* and SNP *l*, associated to 4 genotype combinations of the two SNPs. If the coefficients are known, penetrances for every possible genotype combination could be determined from the logistic models and then used for data generation.

For easier interpretation, the coefficients α, β and γ are not directly used by the GWAsimulator program as the input parameters. Instead, they are calculated within the program from user-specified relative risks (RR). When there is no interaction, a pair of genotypic relative risks specify the marginal genotypic effects at SNP *i*, $r_{i1} = f(g_i = 1) / f(g_i = 0)$ and $r_{i2} = f(g_i = 2) / f(g_i = 0)$, where $f(\cdot)$ is the penetrance function. When pair-wise interactions exist, in addition to the RRs at each locus, departure of relative risks for genotype combinations from the product of corresponding marginal relative risks is also specified. For interaction between SNP *k* and SNP *l*, departure from product of marginal relative risks is defined as

$$d_{kl,uv}=\frac{f(g_k=u,g_l=v)/f(g_k=0,g_l=0)}{r_{ku}r_{lv}}.$$

If there is no interaction, the SNP effects are independent. Thus, instead of using formula (4), coefficients β could be estimated one by one using a set of single-locus models

$$\text{logit}(f(g_i))=\text{logit}(\Pr(\text{affected}|g_i))=\alpha_i+\beta_{i1}I_{\{g_i=1\}}+\beta_{i2}I_{\{g_i=2\}} \quad (6)$$

In case when there are pair-wise interactions, similar simplification could also be used if a pair of interacting SNPs could be taken as an independent group of factors from other SNPs. In this case, we have

$$\begin{aligned}&\text{logit}(\Pr(\text{affected}|g_k,g_l))\\&\quad=\alpha_{kl}+\beta_{k1}I_{\{g_k=1\}}+\beta_{k2}I_{\{g_k=2\}}+\beta_{l1}I_{\{g_l=1\}}+\beta_{l2}I_{\{g_l=2\}}+\gamma_{kl,11}I_{\{g_k=1,g_l=1\}}\\&\quad+\gamma_{kl,12}I_{\{g_k=1,g_l=2\}}\\&\quad+\gamma_{kl,21}I_{\{g_k=2,g_l=1\}}\\&\quad+\gamma_{kl,22}I_{\{g_k=2,g_l=2\}}\end{aligned} \quad (7)$$

In GWAsimulator, β coefficients are calculated from (6), regardless of whether interactions are involved. After that, γ coefficients for interaction terms are obtained from the two-locus model (7) with the same β coefficients obtained previously. This simple 2-step estimation is not always appropriate. If there are interaction effects, the estimation of γ from (6) ignoring the interaction terms might not be correct. Besides this, equation (7) does not always hold true, either. A simple example is that when a risk SNP is involved in interactions with multiple SNPs, it is not possible to single out this SNP with a single interacting SNP to estimate the pair-wise interaction coefficients.

A general modeling approach using Penetrance tables in simGWA

In the present work, we provide and evaluate methods to correctly specify multilocus disease models with or without epistatic effects, and with either pairwise or high-order interactions. Since penetrance tables are a more general and flexible way to precisely characterize complex joint effects of multiple risk loci, methods for conversion among different model specification methods are developed using penetrance table as the primary characterization of disease models. All methods described below are implemented in an R package called simGWA that produces the correct multilocus penetrance table, and then simulates disease loci genotypes and applies retrospective sampling by a modified GWAsimulator engine to rapidly generate genome-wide marker genotype data (see Figure 1).

(1) Correctly modeling pair-wise interactions using relative risks

Our first method addresses the misspecification problem of GWAsimulator when interaction exists, by correctly calculates the logistic model coefficients from given values of relative

risks (RR). Because direct estimation requires solving high-dimensional nonlinear functions that can quickly become intractable, we developed an iterative approach to circumvent the problem. First, the departure from product of marginal RRs is converted to departure from the product of 2 "boundary" joint RRs involving reference genotypes at individual disease locus (see definition of $d'_{kl,uv}$ below). Then, values of the latter are used in iterative numerical computation to obtain the logistic model coefficients and the full penetrance table. This procedure is Option 1 shown in Figure 1 as path from (A1) to (A2) and then to (B).

Marginal RR and joint RR—Marginal RR is commonly used when describing the effects of individual SNPs. However, when it comes to interactions, joint RR (the risk ratio of a joint genotype ($g_k = u$, $g_l = v$) to the reference ($g_k = 0$, $g_l = 0$) : $RR_{kl,uv} = (g_k = u, g_l = v) / (g_k = 0, g_l = 0)$) gives a better characterization of the relative risks comparing genotype combinations.

For two disease loci k and l, the departure from product of marginal relative risk of having genotype with u copies of disease allele at k and v copies at l is defined as the ratio of the joint RR to the product of the 2 marginal RRs:

$$d_{kl,uv}=\frac{RR_{kl,uv}}{r_{ku}r_{lv}}=\frac{f(g_k=u,g_l=v)/f(g_k=0,g_l=0)}{\{f(g_k=u)/f(g_k=0)\}\{f(g_l=v)/f(g_l=0)\}}$$

while the departure from product of 2 "boundary" joint RR is defined as

$$d'_{kl,uv}=\frac{RR_{kl,uv}}{RR_{kl,u0}\cdot RR_{kl,0v}}=\frac{f(g_k=u,g_l=v)/f(g_k=0,g_l=0)}{\{f(g_k=u,g_l=0)/f(g_k=0,g_l=0)\}\{f(g_k=0,g_l=v)/f(g_k=0,g_l=0)\}}.$$

For a given set of values of $d_{kl,uv}$, the values of $d'_{kl,uv}$ can be determined by solving a system of linear equations.

Numerical algorithm to estimate coefficients and penetrance table—Assuming logistic model in formula (5), the coefficients need to estimate are α, β_{i1} and α_{i2} ($1 \le i \le m$), and for all interacting SNP pairs between SNPs k and l, $\gamma_{kl,11}$, $\gamma_{kl,12}$, $\gamma_{kl,21}$ and $\gamma_{kl,22}$. The estimation is achieved by iterative calculating the coefficients and constructing the penetrance table. At each iteration, we first construct the full penetrance table from the current set of coefficients. From the full penetrance table, the joint RR and marginal RR values are calculated and compared with those specified originally in the model. The difference between the current RR values to those from the model specifications is used to update the logistic model coefficients accordingly to reduce the difference. After many iterations, the full penetrance table conforms well to all joint RR and marginal RR constraints with neglectable bias.

We start by letting $\alpha^{(0)} = \log(K / (1-K))$ and all SNP coefficients $\beta_{iu}^{(0)}$ and $\gamma_{kl,uv}^{(0)}$ equal to 0. The penetrance for each genotype is then a constant, $\Pr(\text{affected} \mid g_1,\ldots g_m) = K$.

Suppose at iteration *s*, the previously estimated coefficients are $\alpha^{(s-1)}$, $\beta_{iu}^{(s-1)}(1 \le i \le m, u \in \{1,2\})$, $\gamma_{kl,uv}^{(s-1)}$ (*k*,*l* are interacting SNPs, *u*,*v*)∈{1,2}), and the penetrance table calculated from the coefficients is $f^{(s-1)}$. The following steps are taken to update their values in the iteration.

Step 1: for each 1≤*i*≤*m*, update β_{iu} and penetrance table *f*. From the full penetrance table $f^{(s-1)}$, the marginal penetrances at locus *i* are calculated by weighed average of all penetrances involving a certain genotype at this locus, and the weight is the corresponding multilocus genotype frequency. Denote the derived marginal penetrances as $f_{i0}^{(s-1)} f_{i1}^{(s-1)}$, and $f_{i2}^{(s-1)}$ for the 3 genotypes at SNP *i*. Then update $\beta_{iu}^{(s)} = \beta_{iu}^{(s-1)} + \log(\frac{r_{iu}}{f_{iu}^{(s-1)} / f_{i0}^{(s-1)}})$, and update the penetrance table based on the new set of coefficients. Denote the final penetrance table after updating $\beta_{iu}^{(s)}$ for all SNPs as $f^{(s,1)}$

Step 2: update the value of *a*. After step 1, the disease prevalence might not equal to *K*. From penetrance table $f^{(s,1)}$, calculate the current disease prevalence $K^{(s,1)}$, then *a* is updated to $\alpha^{(s,2)} = \alpha^{(s-1)} + \log(K / K^{(s,1)})$. Accordingly, the new penetrance table from the coefficients is now $f^{(s,2)}$.

Step 3: for each pair of interacting SNPs *K* and *l*, update the values of $\gamma_{kl,uv}(u,v = 1,2)$. In this step we estimate departure from joint relative risks from $f^{(s,2)}$. Estimations are $d'_{kl,uv}{}^{(s,2)}$. Then the updated γ values are $\gamma_{kl,uv}{}^{(s)} = \gamma_{kl,uv}{}^{(s-1)} + \log(d'_{kl,uv} / d'_{kl,uv}{}^{(s,2)})$. Denote the penetrance table after updating for all $\gamma_{kl,uv}$ as $f^{(s,3)}$

Step 4: update the value of *a* as in step 2. After this updating, the logistic model coefficients are now $\alpha^{(s)}$, $\beta_{iu}{}^{(s)}$, $\gamma_{kl,uv}{}^{(s)}$ and the penetrance table calculated from them is $f^{(s)}$.

Step 5: check the maximum change of values to all coefficients in step 1–4 against a pre-set tolerance threshold (default value of 10^{-10} is used for results shown below). Iterations of step 1–4 are repeated until the maximum change is below tolerance or a maximum number of iterations is reached.

(2) Effectively modeling higher-order interactions

High-order interactions can be modeled in simGWA either by logistic models (this is Option 2 marked in Figure 1 as (B)) or by sampling penetrance tables generated using the previously developed simP R package[Yang and Gu 2008] (this is Option 3 marked in Figure 1 as (C)).

For the logistic model approach (Option 2), a formula similar to equation (4) is used to determine the penetrances of multi-locus genotypes, with additional higher-order product terms for interactions among multiple risk SNPs. Users need to specify all coefficients in the model. The simGWA package automatically calculates the penetrance table when interactions are limited between pairs of SNPs. Going beyond pair-wise interactions, the users have to calculate and specify each penetrance values. Although the calculation is straightforward, we discourage the use of this approach when modeling higher than pairwise interactions because the biologically interpretation of the higher-order product terms

becomes less clear. Instead, we recommend directly assign multilocus penetrances conforming to assigned marginal effects. This can be done by simP[Yang and Gu 2008] (Option 3), a previously developed R package that can perform two very useful functions: First, it can generate unlimited number of random penetrance tables that satisfy a given set of marginal relative risk constraints. Second, for any given penetrance table, it quickly evaluates the effects of single SNPs, collective effects of interactions, and the fraction of disease variation explained by the corresponding genetic model. This information could aid selecting interesting interaction models for data simulation. For example, using simP, we were able to generate hundreds of genetic models with null marginal effects for all risk SNPs, but their joint effects account for a substantial amount of disease variability.

(3) Special penetrance tables prescribing known biological models

Some well-known interaction models can be directly specified using penetrance tables (this is Option 4 marked in Figure 1 as (D)). Below are two such examples.

1. Heterozygous model. Occurrence of any risk genotype from different loci causes the disease. In Table 1, occurrence of AA genotype or any B allele causes the disease (penetrance is 1).
2. Threshold model. Disease phenotype manifests (penetrance is 1) when the total number of risk alleles/genotypes reaches a threshold.

Evaluation of simGWA performances

To evaluate the performances of simGWA, we applied the simulator over a range of genetic models of multilocus interactions and used HapMap phased data for Caucasians (CEU) as template. The simulated GWAS datasets include a total of 676,565 SNPs on the Affymetrix Genome-Wide Human SNP Array 6.0 to mimic the real genotyping platform.

The GWAS data were simulated for a binary trait with 5 risk SNPs. They locate on 5 randomly selected chromosomes. Among the five SNPs, SNP1, SNP3 and SNP4 have no marginal effect at all; SNP2 has a multiplicative effect, and the relative risks of genotypes with one and two copies of risk alleles (compared with that of no copy) are 1.5 and 2.25, respectively; SNP5 has a dominant effect, and the relative risk of both risk genotypes is 2. Namely, $r_{11} = r_{12} = 1$; $r_{21} = 1.5$, $r_{22} = 2.25$; $r_{31} = r_{32} = 1$; $r_{41} = r_{42} = 1$; $r_{51} = r_{52} = 2$.

Assuming the marginal effects described above, two types of models were considered in terms of SNP-SNP interactions: one with no interaction at all, and the other with pairwise interactions between 2 SNP pairs (SNP1 interacting with SNP2, SNP3 with SNP4). In the latter, the effect sizes of the pairwise interactions, as measured by departure from product of marginal RRs, are $d_{12,11} = 0.2$, $d_{12,12} = 1$, $d_{12,21} = 2$, $d_{12,22} = 3$, $d_{34,11} = 0.5$, $d_{34,12} = 2$, $d_{34,21} = 2$, $d_{34,22} = 1.6$. For each model, we generated data sets containing 676,565 SNPs for 2000 cases and 2000 controls using simGWA, by: (1) converting the marginal RR model to joint RR model; (2) numerically calculating logistic model coefficients and generating the penetrance table; (3) generating genetic data. To compare performance, we also generated datasets using GWAsimulator and the same parameters for disease models with or without

pairwise interactions. Under each of the 2 models, 1,000 replication datasets were generated by simGWA and GWAsimulator, respectively.

Results

Comparable performance in terms of Local LD structure & computational time

Both programs took about 52 minutes to simulate genotypes of 676,565 SNPs for 4000 subjects with a single thread on a Linux machine with 2 CPUs of Intel Xeon 5430 Quad Core 2.66 GHz and 32 GB of memory.

For both simGWA and GWAsimulator, genome-wide LD structures in the simulated data were comparable to those in the real population of HapMap CEU. A typical example is given in Figure 2, which confirms that local LD structures by HaploView [Barrett, et al. 2005] were faithfully maintained in both datasets generated by the two simulators.

Agreement between simGWA and the GWAsimulator when there is no interaction

The two simulators handle parameters for disease models differently. However, we may compare effect sizes of each risk SNP estimated based on the simulated penetrance tables. Table 2 summarizes the comparisons when interactions exist or not, expressed in terms of genetic information loss (the reduction in explained heritability) if an individual SNP was ignored. When there is absolutely no interaction (“pure marginal effect” model), marginal effect size of individual risk SNPs were almost identical by both simulators.

For every pair of datasets generated by the two simulators, difference in genotype distributions was tested by χ^2; at each risk SNP, in cases and in controls separately. There were 5000 tests comparing the genotypes at 5 risk SNPs in 1,000 cases datasets, and another 5000 tests in 1,000 controls datasets. The smallest p-value from the 10000 tests was 0.070. This confirms that the distributions of risk genotypes were not different in datasets generated by the 2 simulators. Further tests were carried out comparing distributions of all two-locus combined genotypes; the similarity still holds.

simGWA correctly simulates pair-wise interactions

When there were interactions, simGWA correctly calculated all interaction terms in the penetrance table, which could differ substantially from those used by GWAsimulator even though the disease model is specified in the same manner. This is clearly seen in the bottom half of Table 2: there were differences both in the total heritabilities calculated by the two methods, and in the effect sizes of the risk SNPs (measured as the decrease in explained disease variation when a risk SNP was ignored). The differences were due to overly simplified specification of the logistic model coefficients in GWAsimulator for SNPs involved in interactions, as demonstrated in Table 3. If two SNPs have no marginal effect nor any interaction between them, their combined effects should be null, such as in the case of SNP1 and SNP3, or SNP1 and SNP4.

However, as shown in Table 3, while the penetrance table calculated by simGWA resulted in a correct value of 0 for the combined effects of 2 such pairs of SNPs, substantial nonzero values (0.13 for both pairs) were assigned by GWAsimulator. This is supported by

association test results on the simulated datasets. Single-SNP association test p values should approximately follow the uniform distribution when the SNP has no marginal effect. As seen in Figure 3, under models of no interaction, distributions of single-SNP tests for SNP1, SNP3, and SNP4 in both simGWA- and GWAsimulator-generated datasets follow perfectly the uniform distribution (Panel A); however, under interaction models, the distributions in GWAsimulator-generated datasets completely diverged from the uniform (Panel B) even though these SNPs had no marginal effects. Similar observations were made for interaction tests and displayed in Figure 4. Again, under pure marginal effect models, p values of pairwise interaction tests for SNP1-SNP3 and SNP1-SNP4 correctly follow the uniform distribution in both simGWA- and GWAsimulator-generated datasets. But when there were interactions, the distributions in GWAsimulator-generated datasets completely diverged from the uniform (Panel B) even though no interaction effects were simulated for these SNP pairs.

Discussion

We presented a novel method for correctly specifying SNP interaction effects and an improved GWAS data simulator using this method called simGWA. Penetrance table is used as the fundamental characterization of disease models, and commonly-used means for interaction model specification (deviation from product of relative risks or logistic model coefficients) are converted to use correct penetrance tables. Arbitrary logistic models or a general-purpose penetrance generator (simP) were used to generate penetrance tables for high-order interactions. Genotype simulation in simGWA is built on the highly efficient GWAsimulator[Li and Li 2008]. Before GWAsimulator, many used the coalescent model[Donnelly and Tavare 1995; Hudson 2002] of population genetics or forward time simulation[Peng and Amos 2010; Pinelli, et al. 2012] to reconstruct the evolutionary history. Although the approach works well for sampling a theoretical population that follows the Wright–Fisher model[Hudson 2002], the simulators are generally not as efficient for GWAS data simulation. Moreover, GWAsimulator adopts an empirical approach and the “retrospective sampling” based on real-population templates, and works excellently when there are no interactions. simGWA fully takes advantage of its efficient simulation engine and by employing new methods to correctly specify SNP interaction effects. This resulted in a useful tool for rapid generation of GWAS data under complex interaction models for studying complex disease.

Existing methods such as GENS[Amato, et al. 2010] and GENS2[Pinelli, et al. 2012] also used multi-locus penetrance tables to model gene-environment interactions. These methods were limited to GxE interactions involving at most 2 disease loci and a single environment factor. It would be interesting to see if the methods can be combined with that of simGWA to simulate GxE interactions involving more environment variables and higher-order penetrance tables.

Although the present work is focused on simulating GWAS data for binary traits, it is possible to extend simGWA to simulate GWAS data for quantitative traits for studies using population-based sampling. However, correct modeling of higher-order interactions for

sampling based on quantitative trait values will not be straightforward and deserves further investigation.

We note that simGWA has some limitations similar to GWAsimulator. For example, flaws in the template phase data (e.g., ascertainment bias) would be passed to the generated data. Also, long-range LDs are not considered; and it allows only one disease locus on each chromosome. These flaws may be remediable in many situations. For example, if there is need to simulate multiple disease loci on the same chromosome in different LD blocks, one can simulate that chromosome in multiple chunks each harboring a risk SNP, with possibly slight loss of LD information at the ends connecting the chunks.

In summary, simGWA provides a rapid GWAS data simulator that is able to mimic realistic LD and correctly model complex interactions among risk SNPs. As more and more efforts are put to in-depth analysis of GWAS data to find “missing heritability”, many sophisticated analytical methods are in development and we anticipate that simGWA will provide a useful tool for method evaluation.

Acknowledgments

This research was supported in part by NIH grants HL091028 and HL071782, DA027995, and an AHA grant 0855626G.

References

Amato R, Pinelli M, D'Andrea D, Miele G, Nicodemi M, Raiconi G, Cocozza S. A novel approach to simulate gene-environment interactions in complex diseases. BMC bioinformatics. 2010; 11(1):8. [PubMed: 20051127]

Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, Chatterjee N, Kraft P, Van Steen K. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Human Genetics. 2012; 131(10):1591–1613. [PubMed: 22760307]

Barrett J, Fry B, Maller J, Daly M. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21(2):263–265. [PubMed: 15297300]

Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics. 2009; 10(6):392–404.

Donnelly P, Tavare S. Coalescents and genealogical structure under neutrality. Annual review of genetics. 1995; 29(1):401–421.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. American journal of human genetics. 2004; 75(1):35. [PubMed: 15148658]

Gao H, Wu Y, Li J, Li H, Li J, Yang R. Forward LASSO analysis for high-order interactions in genome-wide association study Briefings in Bioinformatics. 2013

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, Huang W, Liu B, Shen Y. The international HapMap project. Nature. 2003; 426(6968):789–796. [PubMed: 14685227]

Gyenesei A, Moody J, Laiho A, Semple CA, Haley CS, Wei W-H. BiForce Toolbox: powerful high-throughput computational analysis of gene– gene interactions in genome-wide association studies. Nucleic acids research. 2012; 40(W1):W628–W632. [PubMed: 22689639]

Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. Bioinformatics. 2003; 19(3):376–382. [PubMed: 12584123]

Hindorff, L.; MacArthur, J.; Morales, J.; Junkins, H.; Hall, P.; Klemm, A.; Manolio, T. [Acessed Aug 6] A Catalog of Published Genome-Wide Association Studies. 2013. Available at: http://www.genome.gov/gwastudies.

Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002; 18(2):337–338. [PubMed: 11847089]

Jin L, Zhu W, Guo J. Genome-wide association studies using haplotype clustering with a new haplotype similarity. Genetic epidemiology. 2010; 34(6):633–641. [PubMed: 20718046]

Kirino Y, Bertsias G, Ishigatsubo Y, Mizuki N, Tugal-Tutkun I, Seyahi E, Ozyazgan Y, Sacli FS, Erer B, Inoko H. Genome-wide association analysis identifies new susceptibility loci for Behcet's disease and epistasis between HLA-B [ast] 51 and ERAP1. Nature genetics. 2013; 45(2):202–207. [PubMed: 23291587]

Li C, Li M. GWAsimulator: a rapid whole-genome simulation program. Bioinformatics. 2008; 24(1): 140–142. [PubMed: 18006546]

Lucas G, Lluís-Ganella C, Subirana I, Musameh MD, Gonzalez JR, Nelson CP, Sentí M, Schwartz SM, Siscovick D, O'Donnell CJ. Hypothesis-based analysis of gene-gene interactions and risk of myocardial infarction. PloS one. 2012; 7(8):e41730. [PubMed: 22876292]

Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A. Knowledge-driven analysis identifies a gene–gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. PLoS genetics. 2012; 8(5):e1002714. [PubMed: 22654671]

MacLellan WR, Wang Y, Lusis AJ. Systems-based approaches to cardiovascular disease. Nature Reviews Cardiology. 2012; 9(3):172–184.

Manolio TA. Bringing genome-wide association findings into clinical use. Nature Reviews Genetics. 2013; 14(8):549–558.

O'Seaghdha CM, Fox CS. Genome-wide association studies of chronic kidney disease: what have we learned? Nature Reviews Nephrology. 2011; 8(2):89–99.

Oh S, Lee J, Kwon M-S, Weir B, Ha K, Park T. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. BMC Bioinformatics. 2012; 13(Suppl 9):S5. [PubMed: 22901090]

Pandey A, Davis N, White B, Pajewski N, Savitz J, Drevets W, McKinney B. Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder. Translational psychiatry. 2012; 2(8):e154. [PubMed: 22892719]

Peng B, Amos CI. Forward-time simulation of realistic samples for genome-wide association studies. BMC bioinformatics. 2010; 11(1):442. [PubMed: 20809983]

Pinelli M, Scala G, Amato R, Cocozza S, Miele G. Simulating gene-gene and gene-environment interactions in complex diseases: Gene-Environment iNteraction Simulator 2. BMC bioinformatics. 2012; 13(1):132. [PubMed: 22698142]

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics. 2010; 87(3):325–340.

Willer CJ, Mohlke KL. Finding genes and variants for lipid levels after genome-wide association analysis. Current Opinion in Lipidology. 2012; 23(2):98–103. [PubMed: 22418572]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics. 2010; 86(6):929–942.

Yang J, Ferreira T, Morris AP, Medland SE, Consortium GIoAT, Consortium DGR, Meta a, Madden PAF, Heath AC, Martin NG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics. 2012; 44(4):369–375. [PubMed: 22426310]

Yang W, de las Fuentes L, Dávila-Román VG, Gu CC. Variable set enrichment analysis in genome-wide association studies. European Journal of Human Genetics. 2011; 19(8):893–900. [PubMed: 21427759]

Yang W, Gu CC. A Characterization of the Parameter Space for Highorder Epistasis. Genetic Epidemiology. 2008; 32:722.

Yang W, Gu CC. Random forest fishing: a novel approach to identifying organic group of risk factors in genome-wide association studies. European Journal of Human Genetics. 2013

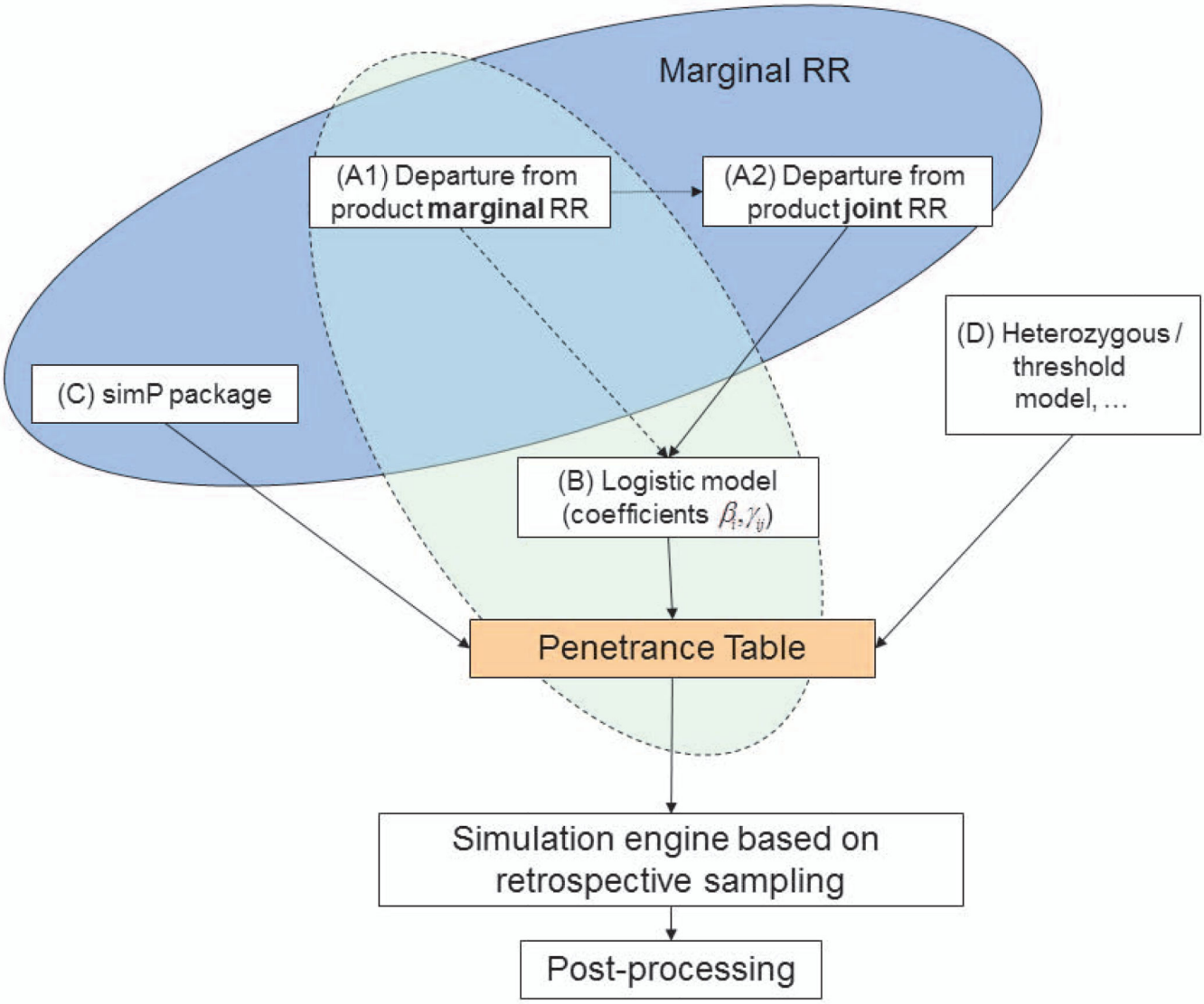

Figure 1.

Schematic representation of four approaches for specifying disease models and methods for converting them to penetrance tables used by simGWA. The models used in GWAsimulator are shown in the dashed oval, in which, marginal relative risks and departure from product of marginal RR (A1) are required from users to estimate logistic model coefficients (B), and then the penetrance table is created internally in GWAsimulator. We identified problem in estimation of logistic model coefficients from model (A1) to (B). In simGWA package, we first convert departure from product of marginal RR (A1) to departure from product of joint RR (A2), and then estimate logistic coefficients and penetrances by numerical calculations. Other ways to generate penetrance tables in simGWA include: (B) calculating penetrances directly from logistic model given coefficients values for low-order interactions; (C) using package simP to get joint penetrances for high-order interactions; (D) using other user specified penetrance table structures, such as heterogeneous model and threshold model (see

methods). In methods A1, A2 and C (shown in solid oval), the marginal relative risks are constrained to model the marginal effects at the disease loci.

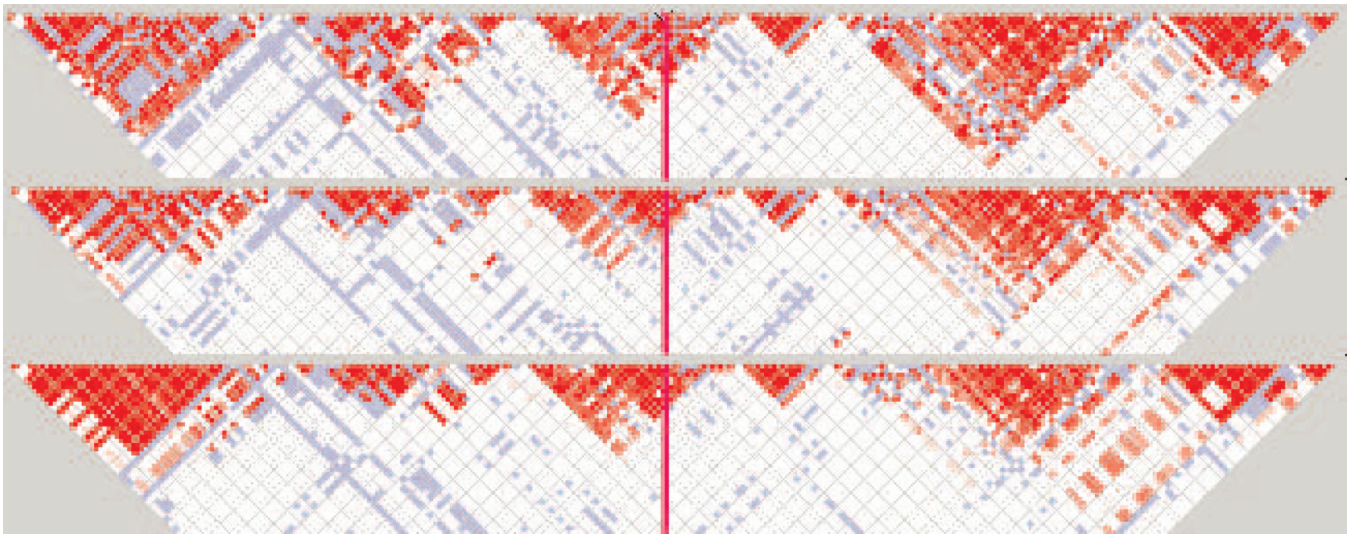

Figure 2. Comparison of LD structures

An example of LD structures in simulated samples in a region of 150 SNPs flanking a disease locus compared to that in the HapMap CEU template (top panel). The bottom 2 panels show results using the GWAsimulator (middle) and simGWA(Bottom). Each data set has 60 individuals. Pink vertical bar shows the position of a risk SNP. The LD structure were plot using Haploview.

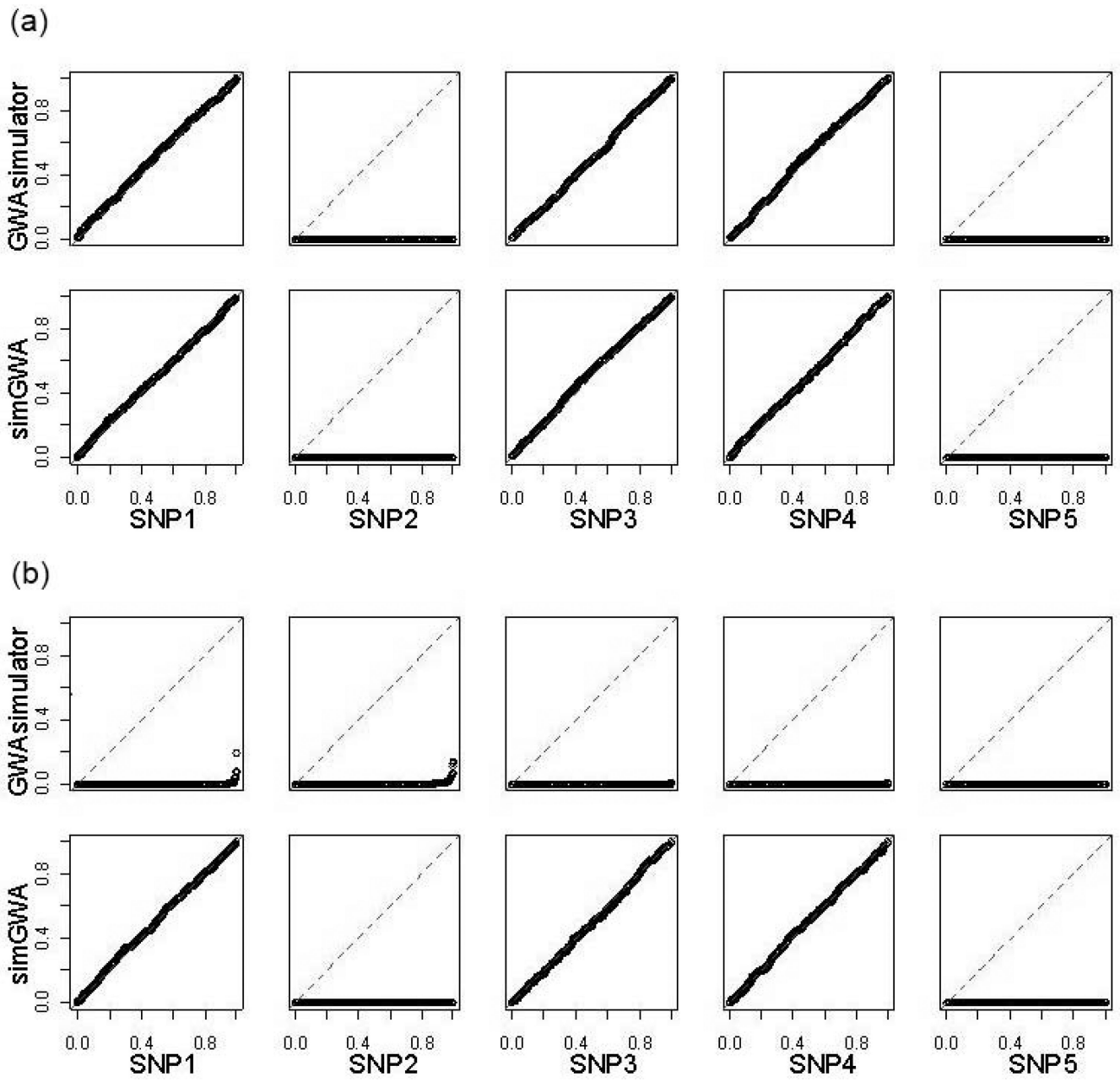

Figure 3. QQ plots compare p-value distributions for χ^2 test for the risk SNPs

Each plot compares the distributions of χ^2 test p values (x-axis) on 1000 simulated data sets against the uniform distribution (y-axis). Close to the diagonal line means the p values are approximately uniform. If there is any signal, the plot lies below the diagonal. The two panels are: (A) Model without interactions; (B) Interaction model.

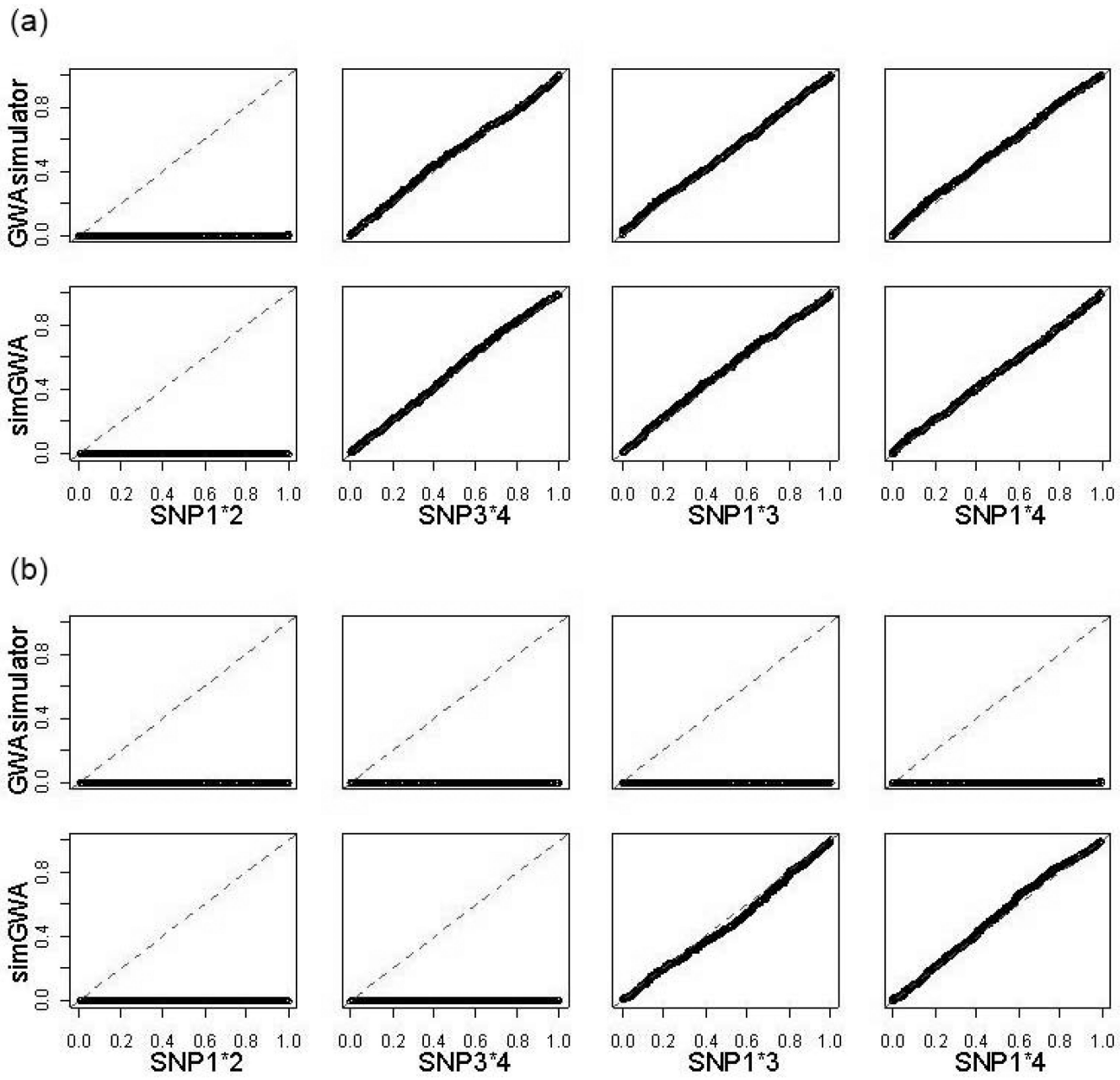

Figure 4. QQ plots compare p-value distributions for interaction test for the risk SNP pairs

Each plot compares the distributions of interaction test p values (x-axis) on 1000 simulated data sets against the uniform distribution (y-axis). Close to the diagonal line means the p values are approximately uniform. If there is any signal, the plot lies below the diagonal. The two panels are: (A) Model without interactions; (B) Interaction model.

Table 1

Penetrance table of a heterozygous model with two SNPs. Occurrence of risk genotype AA or allele B results in a penetrance of 1.

	bb	bB	BB
aa	0	1	1
aA	0	1	1
AA	1	1	1

Table 2

Characteristics of the simulated disease models.

			Information loss by ignoring a SNP				
		h^2	SNP1	SNP2	SNP3	SNP4	SNP5
Pure marginal effect	simGWA	0.04	0.00	0.27	0.00	0.00	0.74
	GWAsimulator	0.04	0.00	0.26	0.00	0.00	0.75
Interaction model	simGWA	0.19	0.54	0.60	0.26	0.26	0.20
	GWAsimulator	0.08	0.31	0.32	0.20	0.20	0.39

Models were built using simGWA and the GWAsimulator when there was no interaction or there were 2 interacting SNP pairs (SNP1 and SNP2, SNP3 and SNP4). Of the 5 SNPs, 3 do not have any marginal effect (SNP1, SNP3, SNP4). The 3rd column in the table shows the proportion of disease variation that is explainable by the joint effect of all 5 SNPs (heritability of disease). The next few columns summarize the genetic information loss by ignoring any of the 5 SNPs

Table 3

Summary of interaction effects when there are interactions in the models.

		SNP1*2	SNP3*4	SNP1*3	SNP1*4
Joint effects of SNP pairs	simGWA	0.55	0.22	0.00	0.00
	GWAsimulator	0.35	0.25	0.13	0.13
Interactions effects of SNP pairs	simGWA	0.49	0.22	0.00	0.00
	GWAsimulator	0.23	0.10	0.00	0.00

Models were built using simGWA and the GWAsimulator when there were 2 pairs of SNPs with pair-wise interactions (SNP1 and SNP2, SNP3 and SNP4). For the two SNP pairs that really interact (SNP1*2, SNP3*4), and two other SNP pairs that entails no interactions (SNP1*3, SNP1*4), the interaction effects in the penetrance tables are summarized. “Information in SNP pairs” shows the proportion of variance in the total heritability that is explainable by only considering the SNP pair (joint effect). The last two rows of the tables show the variance explainable by the pair-wise interaction of the two SNPs.