DOI 10.1007/s10038-007-0140-7

Prediction of complex traits based on the epistasis of multiple haplotypes

Kung-Hao Liang · Ying-Jye Wu

Received: 21 November 2006 / Accepted: 7 March 2007 / Published online: 11 April 2007

Abstract Analysis of epistasis, or gene–gene interactions, is of particular importance for revealing the molecular mechanisms of complex human diseases. Multiple genes, each of which has a moderate effect, might interact and produce a complex phenotypic trait. In this paper, we present a novel method of epistasis analysis, utilizing multiple phase-resolved haplotypes residing in different genomic regions. Prediction models can then be derived from the epistasis to indicate the susceptibility of a person to a dichrotomous phenotypic trait. The simulation results showed that the prediction accuracy of this method is dependent on the penetrance rate of the underlying model. The computation cost, on the other hand, is dependent on the number of genomic regions involved for the complex phenotypic trait.

Keywords Epistasis · Haplotype · Association · Boolean algebra · Complex phenotypic trait · Multiple genomic regions

Introduction

Population-based association study is one of the most important approaches for discovering disease–genotype relationships (Freimer and Sabatti 2004). The disease susceptibility of an individual may be predicted once the disease–genotype relationship is found. Genetic associations have been performed on either unphased single nucleotide polymorphisms (SNPs), or phase-resolved haplotypes (Schaid 2004). A haplotype block spans a chromosomal region where the allelic variants are tightly linked to one another [i.e., in linkage disequilibrium (LD)] (Schaid 2005). A haplotype is a combination of allelic variants which are located within a haplotype block and along a single chromosome (Epstein and Satten 2003). Lengths of typical human haplotype blocks range from a few kilo-bases to several hundred kilo-bases (Meng et al. 2003). The average length of human genes is 27 kb (Carlson et al. 2004), which is approximately at the same scale of haplotype blocks. Haplotype-based associations could detect chromosomal regions which harbor disease-causing variants, even when the variants themselves are not genotyped (Evans et al. 2004; Fallin et al. 2001). In addition, haplotypes are more polymorphic than SNPs, offering a more flexible stratification of the population. Haplotype-based associations have been employed in many disease association studies (Epstein and Satten 2003).

The International HapMap project accomplished a valuable reference of haplotype blocks of the human genome (The International HapMap Consortium 2003). Haplotypes of the entire block can be represented by a smaller set of SNPs referred to as tagging SNPs (Meng et al. 2003). It has been demonstrated empirically that uncommon polymorphisms of drug-related genes can be well represented by haplotypes constructed using tagging SNPs (Kamatani et al. 2004). Therefore, a proper selection of tagging SNPs can reduce the cost, efforts and complexity of the study while maintaining statistical power (Carlson et al. 2004; Goldstein and Cavalleri 2005; Meng et al. 2003). Haplotypes of each individual can be derived from unphased SNPs using a variety of well-tested algorithms such as PHASE (Stephens et al. 2001) or HAPLOTYPER (Niu et al. 2002), among others. The derived haplotypes are then utilized for haplotype-based associations (Fallin et al. 2001).

K.-H. Liang (✉) · Y.-J. Wu
Vita Genomics, Inc., 7F, No.6, Sec.1, Jungshing Rd.,
Wugu Shiang, Taipei County 248, Taiwan
e-mail: kunghao.liang@vitagenomics.com

A complex trait is unlikely to associate prominently with a single allelic variant. On the contrary, it could be the consequence of complex biological mechanisms involving multiple genes in multiple genomic regions. Analysis of epistasis has been advocated for deciphering the complex mechanisms, particularly when each involved gene only demonstrates a minor marginal effect (Bell et al. 2006; Carlborg and Haley 2004). Interaction-based strategy has been demonstrated to outperform locus-by-locus search methods for complex traits (Marchini et al. 2005). Recently, an interaction-based method, GABA, has been proposed for detecting the epistasis among unphased SNPs (Liang et al. 2006). GABA has also been employed on the research of diabetic nephropathy (Hsieh et al. 2006). In this paper, we address the issue of epistasis among haplotypes in multiple genomic regions. The proposed methodology, referred to as the Haplotype Association based on Boolean Algebra (HABA), is an extension of GABA, aiming to overcome the challenges incurred on the epistasis of haplotypes. HABA can be used in conjunction with GABA, as well as traditional locus-by-locus methods, for assessing associations from both SNPs and haplotypes.

Materials and methods

Haplotype association based on boolean algebra

The proposed methodology is designed to discover epistatic effects among phased-resolved haplotypes in multiple genomic regions. The epistatic effects are shown as prediction models to indicate the susceptibility of a person to a dichotomous phenotypic trait.

Denote L as the number of genomic regions. Linkage equilibrium is assumed among regions. Each region accommodates a multi-allelic haplotype profile H_l $\{h_{l,0}$ $,h_{l,1},\ h_{l,2},\ldots h_{l,i}\}$, where i is the index of haplotypes at a particular genomic site l, $0 \leq l < L$. When x bi-allelic SNPs occur within a region, the haplotype profile may have 2^x different haplotypes. The number of real haplotypes in H_l, however, is generally fewer.

Denote $B_l = \{b_{ln1}, b_{ln2} \mid 0 \leq n < N, b_{ln1}, b_{ln2} \in H_l\}$ as the union of pairs of haplotypes at site l carried by the nth individual, where N is the total number of individuals. Denote $T = \{B_l \mid 0 \leq l < L\}$ as the entire dataset, including both case and control individuals. In practice, T is formatted as a two-dimensional table, with each rank representing an individual, and each column representing B_l. Each cell of T corresponds to a pair of haplotypes $\{b_{ln1}, b_{ln2}\}$ of a particular patient n at a particular site l.

A prediction model M comprises a chain of haplotype markers (m_k) joined together by the Boolean operators, multiplication and addition {*, +}. It is denoted succinctly as $M(m_k \mid 0 \leq k < K, K \leq L)$, where K is defined as the number of haplotype markers in the model. Each haplotype marker defines the assessment of a single genomic region B_l, the result of which is either true or false. A haplotype marker is denoted succinctly using a binary-valued vector: $m_k = l\ \langle h_{l,0}, h_{l,1}\ ,h_{l,2},\ldots h_{l,\mathrm{i}}\rangle$. For example, $m_k = l\ \langle 0, 0, 1, 0, 1, 0\rangle$ defines the following assessment on the nth individual:

$$m_k = \begin{cases} \text{True} & \text{if } \textbf{any} \text{ of } b_{\mathrm{ln}\,1} \text{ or } b_{\mathrm{ln}\,2} \text{ equals to } h_{l,2} \text{ or } h_{l,4} \\ \text{False} & \text{otherwise} \end{cases} \tag{1}$$

The complement marker of m_k, denoted as m_k^C, is defined as

$$m_k^C = \begin{cases} \text{True} & \text{if } \textbf{both } b_{\mathrm{ln}\,1} \text{ and } b_{\mathrm{ln}\,2} \text{ equals to } h_{l,0}\ h_{l,1}\ h_{l,2}\ h_{l,3} \text{ or } h_{l,5}; \\ \text{False} & \text{otherwise} \end{cases} \tag{2}$$

The intersection of a marker (m_k) and its complement marker (m_k^C) is an empty set. Their union contains all possible haplotypes in H_l. In this way, H_l is partitioned into two mutually exclusive groups defined by m_k and m_k^C, respectively. One group of haplotypes is associated to the disease susceptibility while the other to non-susceptibility. A typical type of haplotype-based association is performed on each specific haplotype $h_{l,i}$, reflecting the differences of haplotype frequencies between case and control groups. In comparison, the omnibus haplotype-profile test gives an assessment on the entire haplotype frequency profiles of H_l and reports an overall P-value (Fallin et al. 2001). Our method, on the other hand, finds the optimum partition of H_l $\{h_{l,0},\ h_{l,1},\ h_{l,2},\ \ldots\ h_{l,i}\}$, which is a valuable additional information to the above two methods.

A Boolean variable indicates the result of assessment of a haplotype marker. The variables are linked together by Boolean operators to construct M, a Boolean statement. The prediction result of an individual (which is either true or false) is computed from the values of Boolean variables. A Boolean statement can accommodate various types of relationships between variables, including the exclusive OR relationship (see ‘‘Discussion’’).

Model optimization

We aimed to find a model which has the highest prediction performance on the dataset *T*,

$$M_{\text{optimum}} = \arg\max_{M} F(T) \tag{3}$$

where *F*(*T*) is the Fitness score indicating the prediction performance of *M* on *T*. Similar to GABA, HABA adopts the Genetic algorithm for the optimization process (Liang et al. 2006), where candidate models are constantly altered by either mutation or cross-over operations for finding the adequate combinations of haplotype markers. Denote *R* as the case population and R^C the control population, thus $T = R + R^C$. The Sensitivity of a model *M* on *T* is $Pr(M = 1|R)$, the probability of *M* being true within the case population; and the Specificity $Pr(M^C = 1|R^C)$, the probability of M^C being true (*M* being false) within the control population. In this paper, we defined *F* as

$$F = \text{sensitivity} + \text{specificity}. \tag{4}$$

Sensitivity and specificity are two important clinical indexes of prediction performance. In comparison, positive predictive values ($Pr(R|M)$) and negative predictive values ($Pr(R^C|M^C)$) are posterior probabilities which are only adequate when prior probabilities (such as $Pr(R)$) were accurately estimated in the population (Yang et al. 2003). The likelihood ratio ($LR(M)$) is another commonly-used performance index (Yang et al. 2003).

$$LR(M) = \frac{Pr(M|R)}{Pr(M|R^C)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} \tag{5}$$

Simulation

The penetrance rate of the underlying model was an indication of the difficulties of detecting the model accurately. The penetrance rates of a single genetic marker *m* is defined as a conditional probability (Zhao et al. 2003)

$$\text{Penetrance rate of } X = Pr(R|m); \tag{6}$$

The penetrance rate needs to be distinguished from prevalence, the proportion of affected individuals in a population, i.e. $Pr(R)$. We investigated both the penetrance of individual haplotype markers at a single genomic region as well as the penetrance of the entire model involving several haplotype markers $Pr(R|M)$. The aim of HABA is to detect *M* at conditions when individual makers have moderate marginal effects. The sum of $Pr(R|M)$ and $Pr(R^C|M^C)$ has been proved to be greater than 1 (Appendix). To reflect the level of difficulty of the simulation using both $Pr(R|M)$ and $Pr(R^C|M^C)$, we assume they are equal in this simulation, i.e. penetrance rate = $Pr(R|M) = Pr(R^C|M^C)$. Under this assumption, the minimum penetrance rate is 50.

Datasets for the simulation were generated randomly, according to the specified number of cases and controls, as well as *L*. For each region, a multi-allelic haplotype profile H_l $\{h_{l,0},\ h_{l,1}\ ,h_{l,2},\ldots h_{l,i}\}$ was randomly generated. The number of haplotypes *i* was a randomly generated value between 2 and 7, which are commonly observed numbers of haplotypes in the human genome. Each haplotype in H_l was assumed to have equal frequency for simplicity. The underlying models were also generated randomly according to the specified number of markers involved. The markers were randomly chosen from the *L* regions and then randomly determined based on H_l of the simulation dataset. Finally, the haplotypes in the marker regions were randomly modified, meeting the specified penetrance requirement of the underlying model.

Three sets of simulations were conducted. The first set of simulations demonstrated the characteristics/behavior of HABA at various conditions when the underlying models involved various numbers of genomic regions, and had complete and various incomplete penetrance. Twenty different models were generated randomly, one for each condition (number of markers = 1–4; penetrance = 60–100%). The models were then used to dictate the generation of individual genotypes for 1,000 cases and 1,000 controls. Therefore *N* = 2,000. The number of genomic regions (i.e. *L*) was 8. The halting condition of the program was when the best model remained unchanged for 1,000 iterations.

The second set of simulations was designed to evaluate the average performance of HABA on 50 replicated tests. We employed datasets comprising five genomic regions and an embedded model comprising three haplotype markers under a variety of penetrance rates (60–100%). The number of iterations for halting in this test is 300.

The third set of simulations was a permutation test (Hirschhorn and Daly 2005) showing the empirical significance level of the detected model when the dataset contain 100 genomic regions. The number of iterations for halting in this test is 150.

The heuristic parameters of this algorithm were identical to those previously described (Liang et al. 2006). A total of 300 models were used within an iteration of the computation. The Fitness score is defined in Eq. (4). No parsimonious constraints were used, apart from the condition when an equal Fitness score occurs on two candidate models. In such a condition, the one involving fewer markers would be ranked higher.

Table 1 Marginal penetrance of haplotype markers, as well as the marker frequencies, when four genomic regions were involved in the underlying model

Model penetrance (%)	Underlying model (four regions)	Marginal penetrance of haplotype markers		
		Location	Penetrance (%)	Frequency
100	Control			
	1 ⟨1, 0, 1⟩	1	52.72	0.653
	* 4 ⟨1, 0⟩	4	55.06	0.457
	* 2 ⟨1, 0, 0, 1, 0, 0⟩	2	68.69	0.268
	+ 0 ⟨1, 0⟩	0	100	0.287
90	Case			
	3 ⟨1, 0, 1⟩	3	90.52	0.372
	+ 6 ⟨0, 1, 1, 0, 1, 0⟩	6	54.43	0.475
	* 0 ⟨1, 1, 0, 0⟩	0	53.13	0.476
	* 2 ⟨0, 0, 0, 1, 0, 1⟩	2	58.22	0.265
80	Control			
	5 ⟨1, 1, 0, 0⟩	5	54.62	0.492
	* 1 ⟨0, 0, 0, 0, 1, 0, 1⟩	1	66.33	0.271
	+ 7 ⟨0, 1, 0, 0, 0, 1⟩	7	60.56	0.322
	* 4 ⟨0, 1, 0⟩	4	60.86	0.321
70	Control			
	5 ⟨1, 0, 0, 0⟩	5	71.37	0.225
	+ 1 ⟨1, 0, 0, 0⟩	1	58.24	0.235
	* 6 ⟨0, 1⟩	6	51.29	0.484
	* 4 ⟨0, 1⟩	4	51.41	0.478
60	Control			
	0 ⟨1, 0⟩	0	60.69	0.255
	+ 3 ⟨1, 0, 1, 1, 1, 0⟩	3	60.19	0.337
	+ 1 ⟨0, 1⟩	1	51.52	0.388
	* 4 ⟨0, 1⟩	4	52.92	0.339

The penetrance of individual haplotype markers is generally smaller than the penetrance of the entire model

Results

Table 1 presented examples of models which were employed in the first set of simulation. These models all comprise four haplotype markers (each at a particular genomic region) but have various penetrance to the datasets. The marginal haplotype frequency $Pr(m)$ and penetrance $Pr(R|m)$ varies because the models and datasets were randomly generated (Table 1). However, it can be seen that the penetrance of individual haplotype marker was usually smaller than the model penetrance. Thus, the dataset simulates the epistasis where the associations reveal themselves at the combination of multiple haplotype markers, rather than individual markers.

The first set of simulation evaluates the performance of HABA when various numbers of genomics regions (between 1 and 4) were involved, and when various penetrance rates (between 60 and 100%) were observed in the data. The detected models were identical to all the underlying models when they had complete penetrance, resulting in 100% sensitivity and specificity. The underlying models were also detected accurately when their penetrance rates were 90, 80 and 70%. When the penetrance rate of the models was further reduced to 60%, ‘‘over-fitting’’ models were detected instead of accurate underlying models, resulting in Fitness scores higher than the expected value (Table 2). The detected models not only contain some of the correct markers, but also introduce several additional markers (Table 2). From these experiments, we observed that the prediction accuracy is independent of number of regions involved. It is, however, dependent on the penetrance rate of the model.

We also compared the computation cost, shown as the number of iterations, for detecting the optimum model at various conditions (Table 3). The numbers of iteration did not include the additional 1,000 iterations after the optimum models were achieved (see ‘‘Materials and methods’’). Table 3 shows that the average number of iterations is dependent on the number of markers in the underlying model. The penetrance rate, on the other hand, did not affect the computation cost.

Having observed the general performance of the algorithm at a variety of conditions, we conducted the second set of simulations to calculate the average performance,

Table 2 Comparisons of the underlying models and the detected models when the penetrance rate was 60%

No. markers	Underlying model	Detected model	Sensitivity (%)	Specificity (%)	Fitness score (%)
1	Control 0 ⟨0, 0, 0, 1⟩	Control 0 ⟨0, 0, 0, 1⟩ + 2 ⟨0, 0, 0, 0, 0, 1, 0⟩ * 7 ⟨0, 0, 1, 0, 1, 0, 0⟩	53.7	67.6	121.3
2	Control 2 ⟨0, 0, 1⟩ + 7 ⟨1, 0⟩	Control 2 ⟨0, 0, 1⟩ * 2 ⟨1, 0, 0⟩ * 4 ⟨0, 0, 0, 1, 1, 0, 1⟩ + 2 ⟨0, 0, 1⟩ * 5 ⟨0, 0, 0, 1, 0⟩ * 4 ⟨0, 0, 1, 0, 1, 1, 0⟩ + 7 ⟨1, 0⟩	65.1	56.9	122.0
3	Case 4 ⟨0, 1, 1, 0, 1, 0, 0⟩ * 1 ⟨0, 0, 0, 1, 0, 0⟩ + 7 ⟨0, 1, 0, 0, 1, 0⟩	Case 4 ⟨1, 1, 1, 0, 1, 0, 0⟩ * 3 ⟨1, 1, 0, 1, 0, 1⟩ * 0 ⟨1, 1, 0⟩ * 1 ⟨0, 0, 0, 1, 0, 0⟩ + 7 ⟨0, 1, 0, 0, 1, 0⟩	62.1	59.2	121.3
4	Control 0 ⟨1, 0⟩ + 3 ⟨1, 0, 1, 1, 1, 0⟩ + 1 ⟨0, 1⟩ * 4 ⟨0, 1⟩	Control 3 ⟨1, 0, 1, 0, 0, 0⟩ + 3 ⟨0, 0, 0, 1, 1, 0⟩ * 6 ⟨0, 1, 1, 0, 0, 0⟩ * 5 ⟨1, 1, 1, 0, 1, 0⟩ * 1 ⟨1, 0⟩ * 7 ⟨0, 0, 1, 0, 1, 1, 0⟩ + 0 ⟨1, 0⟩ * 7 ⟨0, 1, 1, 0, 1, 0, 1⟩ + 0 ⟨1, 0⟩ * 6 ⟨1, 0, 0, 1, 0, 1⟩	65.7	57.6	123.3

''Over-fitted'' models were detected, resulting in Fitness scores higher than the expected value (i.e. 1.2)

Table 3 The iterations required for finding the optimum model

Iterations	Penetrance rate				
	100%	90%	80%	70%	Average
No. genomic regions involved					
1	6	5	6	4	5
2	60	21	9	72	41
3	771	1,401	1,504	681	1,089
4	629	6,781	2,791	616	2,704

The average number of iteration increased as more genomic regions were involved in the association

including the sensitivity, specificity and computational cost, at various penetrance between 60 and 100%. The underlying model comprises three haplotype markers and characterizes control samples:

$$M = 3\langle 0, 1\rangle + 2\langle 0, 0, 1\rangle * 4\langle 0, 0, 1\rangle.$$

The results are presented in Table 4, where each value is an average of 50 tests. The error sum is defined as the average of the sum of absolute differences between the measured and expected sensitivity and specificity values. The error sum increases as the penetrance decreases, implying that the accuracy depends on the penetrance rate. This is consistent with the observations from the first set of simulation. The error count is defined as the number of tests when the underlying model was not detected among the 50 replicates. Although the underlying model was not always accurately detected, HABA detected approximate models, resulting in small error sum values. The computation cost, measured by the averaged numbers of iterations, does not depend on the penetrance rate.

The third set of simulation is to observe the distribution of performance indexes when the labels of phenotypes (i.e.

Table 4 The averaged performance of 50 tests under various penetrance between 60 and 100%

Penetrance (%)	Sensitivity (%)	Specificity (%)	Error sum (%)	Error count	No. of iterations	SD of no. of iterations
100	100.00	100.00	0.00	0	128.92	105.37
90	90.07	89.20	0.97	6	166.08	164.68
80	80.03	79.86	0.21	1	159.22	123.11
70	70.02	69.99	0.04	1	150.10	124.49
60	60.69	59.14	2.59	13	197.02	178.29

R and R^C) were randomly permuted (Hirschhorn and Daly 2005). An underlying model of two haplotype markers was randomly generated. This model indicates control samples:

$$M = 8\,\langle 1, 0, 1, 0, 0, 0, 0\rangle * 48\,\langle 0, 1, 0\rangle.$$

This model was then used to guide the generation of a dataset, comprising 100 genomic regions, 1,000 cases and 1,000 controls. HABA have repeatedly detected the underlying model twice after 171 and 342 iterations of computation, showing its capability on datasets when $L = 100$. The labels of phenotype were then randomly permuted, resulting in 116 permuted datasets. These datasets simulates the situation where no association occurs, therefore the null distribution can be derived. The histograms of sensitivity and specificity of the models detected by HABA are illustrated in Fig. 1. This shows that the distributions are approximately normal, with the bulk of sensitivity and specificity occurring between 50 and 60%. The histogram of Fitness represents the empirical null distribution of prediction performance, which is illustrated in Fig. 2. The bulk of the null distribution occurs around 110%. Although both sensitivity and specificity have wide distributions, their sum (i.e. the Fitness) values were quite narrowly distributed. According to the null distribution of Fig. 2, the empirical type I error, or the P-value, is smaller than 0.0172 (which is 2/116) if the Fitness of a model is equal or greater than 115%.

Discussion

Analysis of epistasis and haplotype-based association are both important issues for genetic associations. The proposed HABA methodology addresses both issues at the same time. It can be used to construct prediction models involving multiple haplotypes in different genomic regions. HABA enables the discovery of relationship among these haplotypes, facilitating further interpretation on biological mechanisms.

De Morgan duality and mode of inheritance

A model can address both dominant and recessive modes of inheritance at a genomic region. At the level of single haplotype markers, m_k describes the dominant mode of inheritance, while m_k^C accommodates the recessive mode of inheritance (cf. Eqs. 1, 2). According to the De Morgan's law on duality (Liang et al. 2006), if a model is constructed by m_k for predicting either a case or control group with a dominant mode of inheritance, then a corresponding model, consisting of m_k^C, is simultaneously determined for

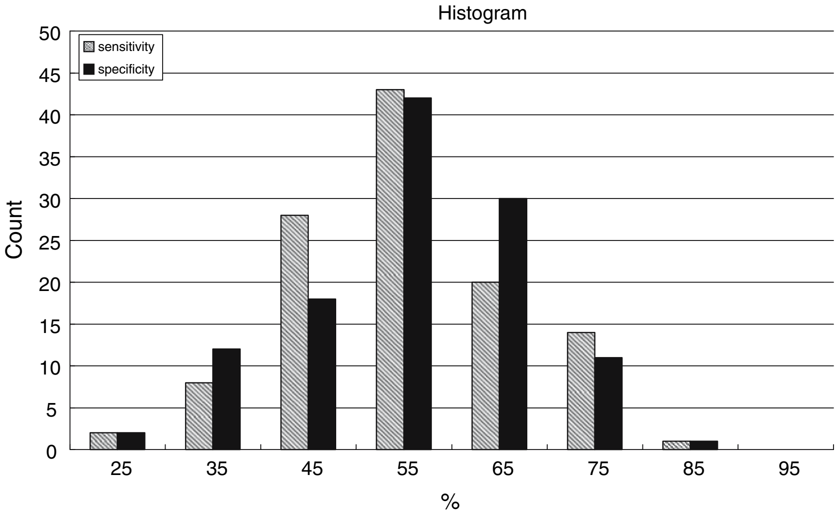

Fig. 1 The histogram of sensitivity and specificity when the label of phenotypes (R and R^C) were permuted randomly

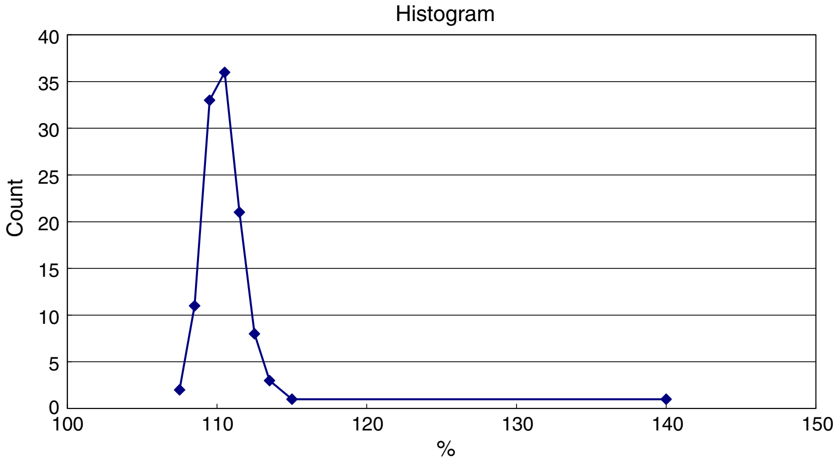

Fig. 2 The histogram of Fitness when the labels of phenotypes (R and R^C) were permuted randomly, simulating the situation where no association occurs to the dataset. Hence, this is an empirical null distribution of performances where the empirical type I error could be estimated

indicating the other group with the recessive mode of inheritance. In other words, $M(m_k \mid 1 \leq k \leq K, K \leq L)$ and $M^C(m_k^C \mid 1 \leq k \leq K, K \leq L)$ are a pair of models for a dichotomous trait, where $Pr(M + M^C) = 1$. Whether M or M^C is associated to case or control depends on model optimization. HABA cannot accommodate the additive mode of inheritance at this moment.

The epistasis between different genomic regions might appear in a more complex format known as the exclusive OR logic (XOR), apart from simple AND/OR relationships. The exclusive OR logic can also achieved by the HABA structure because exclusive OR is equivalent to the following equation composed of {*, +} operators:

$$m_1\ XOR\ m_2 = m_1 * m_2^C + m_1^C * m_2. \tag{7}$$

Therefore, our formation of models can accommodate the XOR relationship between genomic regions.

HABA is mainly designed for datasets where all the haplotypes of each individual have been unambiguously determined. However, it is almost impossible for all the haplotypes in all the regions to be unambiguously determined from phase unknown samples. If only a small portion of missing/ambiguous haplotypes occur among the entire dataset, then the algorithm can temporarily discard those samples with missing/ambiguous haplotypes occurring at the marker site of interest. However, further research on technologies for resolving phases unambiguously, as well as on the improvement of HABA for addressing the ambiguity of haplotypes, is required so as to facilitate the practical use of epistasis analysis on real haplotype datasets.

In conclusion, our simulation results show that this algorithm can detect or approximate the underlying models, provided that the underlying model has reasonably high penetrance (e.g., higher than 70%) in the dataset. The prediction accuracy of this method is dependent on the penetrance rate of the underlying model. The computation cost, on the other hand, is dependent on the number of genomic regions involved for the complex phenotypic trait. This methodology will facilitate the discovery of novel associations based on the epistasis of haplotypes, an important aspect of research on complex diseases.

Appendix: Proof of $Pr(R|X) + Pr(R^C|X^C) > 1$

Making m a maker of a dichrotomous trait R, a sufficiently large relative risk of m against m^C must be observed. That is

$$Pr(R \mid m) > Pr\left(R|m^C\right)$$

Since $Pr(R) + Pr(R^C) = 1$, it can be derived that $Pr(R|m^C) = 1 - Pr(R^C|\mathrm{m}^C)$;

Therefore

$$Pr(R|m) > 1 - Pr\left(R^C|m^C\right)$$

which will give

$$Pr(R|m) + Pr\left(R^C|m^C\right) > 1$$

Hence, $Pr(R|m)$ and $Pr(R^C|m^C)$ cannot be smaller than 0.5 simultaneously.

References

Bell JT, Wallace C, Dobson R, Wiltshire S, Mein C, Pembroke J, Brown M, Clayton D, Samani N, Dominiczak A et al (2006) Two-dimensional genome-scan identifies novel epistatic loci for essential hypertension. Hum Mol Genet 15:1365–1374

Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies. Nature 5:618–625

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120

Epstein MP, Satten GA (2003) Inference of haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329

Evans DM, Cardon LR, Morris AP (2004) Genotype prediction using a dense map of SNPs. Genet Epidemiol 27:375–384

Fallin D, cohen A, Essioux L, Chumakov L, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer' disease. Genome Res 11:143–151

Freimer N, Sabatti C (2004) The use of pedigree, sib-pair and association studies of common disease for genetic mapping and epidemiology. Nat Genet 36:1045–1051

Goldstein DB, Cavalleri GL (2005) Understanding human diversity. Nature 437:1241–1242

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hsieh CH, Liang KH, Hung YR, Huang LC, Pei D, Liao YT, Kuo SW, Bey MSJ, Chen JL, Chen EY (2006) Analysis of epistasis for diabetic nephropathy among Type 2 diabetic patients. Hum Mol Genet 15:2701–2708

Liang KH, Hwang Y, Shao WC, Chen EY (2006) An algorithm for model construction and its applications to pharmacogenomic studies. J Hum Genet 51:751–759

Kamatani N, Sekine A, Kitamoto T, Lida A, Saito S, Kogame A, Lnoue E, Kawamoto M, Harigai M, Nakamura Y (2004) Large scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. Am J Hum Genet 75:190–203

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex disease. Nat Genet 37:413–417

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Schaid DJ (2004) Genetic epidemiology and haplotypes. Genet Epidemiol 27:317–320

Schaid DJ (2005) Power and sample size for testing associations of haplotypes with complex traits. Ann Hum Gene 70:116–130

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

The International HapMap Consortium (2003) The international HapMap project. Nature 426:789–796

Yang Q, Khoury M, Botto L, Friedman J, Dlanders W (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. Am J Hum Genet 72:636–649

Zhao LP, Li SS, Khalid N (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231–1250