รายงานวิจัยฉบับสมบูรณ์

โครงการ: การสร้างแฮปโลไทป์แบบเปลี่ยนความยาวได้สำหรับการศึกษาอันตรกิริยาระหว่างยีน

Variable-Length Haplotype Construction for Gene-Gene Interaction Studies

โดย รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

เมษายน 2552

สัญญาเลขที่ RSA5180006

รายงานวิจัยฉบับสมบูรณ์

โครงการ: การสร้างแฮปโปลไทป์แบบเปลี่ยนความยาวได้สำหรับการศึกษาอันตรกิริยาระหว่างยีน

Variable-Length Haplotype Construction for Gene-Gene Interaction Studies

รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

บทคัดย่อ

รหัสโครงการ: RSA5180006

ชื่อโครงการ: การสร้างแฮปโปลไทป์แบบเปลี่ยนความยาวได้สำหรับการศึกษาอันตรกิริยาระหว่างยีน

ชื่อนักวิจัย: รองศาสตราจารย์ ดร.ณชล ไชยรัตนะ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

E-mail Address: nchl@kmutnb.ac.th

ระยะเวลาโครงการ: 1 ปี

รายงานฉบับนี้ครอบคลุมเทคนิคการจำแนกแบบไม่มีพารามิเตอร์สำหรับการระบุเซตของเครื่องหมายทางพันธุกรรมแบบสองอัลลีลที่สามารถอธิบายความเสี่ยงต่อการเป็นโรคซับซ้อนได้ดีที่สุดในการศึกษาอันตรกิริยาระหว่างยีน เทคนิคที่ได้พัฒนาขึ้นทำงานโดยการสร้างการส่งระหว่างแฮปโปลไทป์ที่ถูกอนุมานจากเครื่องหมายทางพันธุกรรมกับสถานะการเป็นโรคและไม่เป็นโรค เทคนิคจะค้นหาเซตของเครื่องหมายทางพันธุกรรมที่เป็นไปได้ทั้งหมดที่สร้างได้จากเครื่องหมายทางพันธุกรรมที่มี โดยเซตดีที่สุดจะพิจารณาจากความถูกต้องการจำแนกและเกณฑ์การพิจารณาร่วมอีกสองเกณฑ์ ได้แก่ความสามารถในการแพร่จากแฮปโปลไทป์อันดับสูงไปสู่แฮปโปลไทป์อันดับต่ำและอันดับของเซต เนื่องจากมีการสร้างแฮปโปลไทป์แบบเปลี่ยนความยาวได้ในการระบุเซตของเครื่องหมายทางพันธุกรรมดีที่สุด จึงเรียกเทคนิคที่ได้พัฒนาขึ้นว่าเทคนิคการสร้างแฮปโปลไทป์แบบเปลี่ยนความยาวได้สำหรับการศึกษาอันตรกิริยาระหว่างยีน ในการศึกษานี้ได้มีการเปรียบเทียบประสิทธิภาพของเทคนิคที่ได้พัฒนาขึ้นกับเทคนิคการลดมิติหลายปัจจัยและเทคนิคการค้นหาอันตรกิริยาระหว่างแฮปโปลไปท์ในปัญหาอันตรกิริยาสองตำแหน่ง ผลการจำลองแสดงให้เห็นว่าเทคนิคที่ได้พัฒนาขึ้นเหมาะกับทุกสถานการณ์ของอันตรกิริยาที่มีความเชื่อมโยงแบบไม่สมดุลทั้งมากและน้อยระหว่างเครื่องหมายทางพันธุกรรม

คำหลัก: การศึกษาแบบควบคุมกลุ่ม อันตรกิริยาระหว่างยีน แฮปโปลไทป์ ความเชื่อมโยงแบบไม่สมดุล การจำแนกแบบไม่มีพารามิเตอร์

Abstract

Project Code: RSA5180006

Project Title: Variable-Length Haplotype Construction for Gene-Gene Interaction Studies

Investigator: Associate Professor Dr. Nachol Chaiyaratana
King Mongkut's University of Technology North Bangkok

E-mail Address: nchl@kmutnb.ac.th

Project Period: 1 Year

This report presents a non-parametric classification technique for identifying a candidate bi-allelic genetic marker set that best describes disease susceptibility in gene-gene interaction studies. The developed technique functions by creating a mapping between inferred haplotypes and case/control status. The technique cycles through all possible marker combination models generated from the available marker set where the best interaction model is determined from prediction accuracy and two auxiliary criteria including low-to-high order haplotype propagation capability and model parsimony. Since variable-length haplotypes are created during the best model identification, the developed technique is referred to as a variable-length haplotype construction for gene-gene interaction (VarHAP) technique. VarHAP has been benchmarked against a multifactor dimensionality reduction (MDR) program and a haplotype interaction technique embedded in a FAMHAP program in various two-locus interaction problems. The results reveal that VarHAP is suitable for all interaction situations with the presence of weak and strong linkage disequilibrium among genetic markers.

Keywords: Case-control studies, Gene-gene interaction, Haplotype, Linkage disequilibrium, Non-parametric classification

Acknowledgements

I gratefully acknowledge the financial support from the Thailand Research Fund (TRF) through the Research Career Development Grant (Grant Number: RSA5180006).

Nachol Chaiyaratana

April 2009

Executive Summary

This report presents a non-parametric classification technique for identifying a candidate set of single nucleotide polymorphisms (SNPs)—bi-allelic genetic markers—that best describes disease susceptibility in gene-gene interaction studies. The developed technique functions by creating a mapping between inferred haplotypes and case/control status. The technique cycles through all possible marker combination models generated from the available marker set where the best interaction model is determined from prediction accuracy and two auxiliary criteria including low-to-high order haplotype propagation capability and model parsimony. Since variable-length haplotypes are created during the best model identification, the developed technique is referred to as a variable-length haplotype construction for gene-gene interaction (VarHAP) technique. VarHAP has been benchmarked against a multifactor dimensionality reduction (MDR) program and a haplotype interaction technique embedded in a FAMHAP program in various two-locus interaction problems. The interaction scenarios of interests include both epistasis and heterogeneity. The results reveal that VarHAP is suitable for all interaction situations with the presence of weak and strong linkage disequilibrium among genetic markers.

Contents

1. Introduction

Genetic epidemiology is a research field which aims to identify genetic polymorphisms that involve in disease susceptibility. Usual candidate polymorphisms include restriction fragment length polymorphisms (RFLPs), variable numbers of tandem repeats (VNTRs) and single nucleotide polymorphisms (SNPs). In recent years, SNPs are the most common choices due to simplicity and cost reduction in identification protocols. SNPs in diploid organisms are excellent bi-allelic genetic markers for various studies including genetic association, gene-gene interaction and gene-environment interaction. The availability of multiple SNPs on the same gene can also lead to haplotype analysis where genotypes of interest can be phased into pairs of haplotypes.

Traditional techniques for identification of relationship between a single SNP and disease susceptibility status involve various univariate statistical tests including χ^2 and odds ratio tests (Lewis, 2002; Montana, 2006). However, many complementary computational techniques have been developed in the past decade to handle problems that involve multiple SNPs. Heidema et al. (2006) have categorised these multi-locus techniques, which are capable of identifying a candidate SNP set from possible SNPs, into parametric and non-parametric methods. Examples of parametric method cover logistic regression techniques (Nagelkerke et al., 2005) and neural networks (Ritchie et al., 2003). On the other hand, examples of non-parametric method include a set association approach (Hoh et al., 2001), combinatorial techniques (Nelson et al., 2001; Hahn et al., 2003; Culverhouse et al., 2004) and recursive partitioning techniques (Lunetta et al., 2004; Bureau et al., 2005). In some of mentioned parametric (Ritchie et al., 2003; Nagelkerke et al., 2005) and non-parametric (Hahn et al., 2003; Lunetta et al., 2004; Bureau et al., 2005) methods,

pattern recognition and classification approaches have been successfully implemented as their core engines.

In addition to single and multiple SNP analysis, haplotype analysis has also gained attention from genetic epidemiologists. Haplotypes provide a record of evolutionary history more accurately than individual SNPs. Further, haplotypes can capture the patterns of linkage disequilibrium (LD)—a phenomenon where SNPs that are located in close proximity tend to travel together—in genome more accurately. Therefore, haplotypes may enable susceptibility gene identification in complex diseases more effectively than individual SNPs (Silverman, 2007). In lieu of this evidence, haplotype analysis should also be considered in addition to direct genotype analysis. Many computational techniques use haplotypes, which are inferred from multiple SNPs, as problem inputs. For instance, Sham et al. (2004) proposes a logistic regression technique that produces a mapping model between haplotypes and disease status while Becker et al. (2005) combine haplotype explanation probabilities of given genotypes from multiple gene or unlinked region data into a scalar statistic for a univariate test. Nonetheless, haplotypes have rarely been used as inputs for non-parametric classifiers for genetic association and interaction studies.

In this report, a variable-length haplotype construction for gene-gene interaction (VarHAP) technique is proposed. The technique will involve non-parametric classification where haplotypes inferred from multiple SNP data are the classifier inputs. The chosen architecture for non-parametric classifier is the multifactor dimensionality reduction (MDR) technique (Hahn et al., 2003). Similar to the original MDR technique, the proposed technique would be able to identify appropriate candidate SNPs from possible SNPs and can be used in case-control

genetic interaction studies. However, the technique would also be able to handle the situation where disease susceptibility is detectable in different haplotype backgrounds.

2. MDR, Haplotype Inference and Haplotype Explanation Probability

2.1. MDR

MDR is a classifier-based technique that is capable of identifying the best genetic marker combination among possible markers for the separation between case and control samples. Similar to other classification systems, a *k*-fold cross-validation technique provides a means to determine the classification accuracy of the candidate marker model. Basically, the combined case and control samples are randomly divided into *k* folds where $k-1$ folds of samples are used to construct a decision table for the classifier while the remaining fold of samples is used to identify the prediction capability of the constructed decision table. The decision table construction and testing procedure is repeated *k* times. Hence, the samples in each fold will always be utilised both to construct and to test the decision table. The number of cells in a decision table is given by G^{n_c} where n_c is the number of candidate markers selected from possible markers and *G* is the number of possible genotypes according to the marker. For a SNP, which is a bi-allelic marker, *G* is equal to three. During the decision table construction, each cell in the table is filled with case and control samples that have their genotype corresponds to the cell label. The ratio between numbers of case and control samples will provide the decision for each cell whether the corresponding genotype is a disease-predisposing or protective genotype. An example of decision table construction is illustrated in Figure 1. The prediction accuracy of the decision table is subsequently evaluated by counting the numbers of case and control samples in the testing fold that their disease status can be correctly identified using the constructed decision rules. The process of decision table construction and evaluation must be cycled through all or some of possible $2^{n_m}-1$

combinations where n_m is the total number of available markers in the study. The best genetic marker combination is determined from three criteria: prediction accuracy, cross-validation consistency and a sign test *p*-value. Each time that a testing fold is used for prediction accuracy determination, the accuracy of the interested marker combination model can be compared with that from other models that also contain the same number of markers. The model that consistently ranks the first in comparison to other choices with the same amount of markers would have high cross-validation consistency. The non-parametric sign test *p*-value is calculated from the number of testing folds with accuracy greater than or equal to 50%. This single-tailed *p*-value is given by

$$p = \sum_{i=n_a}^{n_f} \binom{n_f}{i} \left(\frac{1}{2}\right)^{n_f} \quad (1)$$

where n_f is the total number of cross-validation folds and n_a is the number of cross-validation folds with testing accuracy ≥ 50% (Hahn et al., 2003). Among three criteria, prediction accuracy is the main criterion for decision making while the other criteria are only used as auxiliary measures. Cross-validation consistency generally confirms that the high rank model can be consistently identified regardless of how the samples are divided for cross-validation. On the other hand, a sign test *p*-value indicates the number of testing folds with acceptable prediction accuracy and hence describes the usability of the model in the classification task. In the situation where two or more models with different number of markers are equally good in terms of prediction accuracy, cross-validation consistency and sign test *p*-value, the most parsimonious model—the combination with the least number of markers—will be the best model.

2.2. Haplotype Inference

With the availability of multiple SNPs from the same gene, haplotypes can be inferred from given genotypes. Let '`0`' and '`1`' denote the major (common) and minor (rare) alleles at a SNP location in a haplotype. A genotype can then be represented by a string, which consists of three characters: '`0`', '`1`' and '`2`'. In the genotype string, '`0`' denotes a homozygous wide-type site, '`1`' denotes a heterozygous site and '`2`' denotes a homozygous variant or homozygous mutant site. A genotype with all homozygous sites or single heterozygous site can always be phased into one pair of haplotypes. On the other hand, a genotype with multiple heterozygous sites can be phased into multiple haplotype pairs. For example, genotype `0102` leads to haplotypes `0001` and `0101` while genotype `0112` leads to two possible haplotype pairs: `0001/0111` and `0011/0101`. Many algorithms exist for haplotype inference (Excoffier and Slatkin, 1995; Stephens et al., 2001; Niu et al., 2002). In this report, an expectation-maximisation algorithm (Excoffier and Slatkin, 1995) is the chosen technique due to its simplicity and implementation efficacy. Regardless of the inference technique employed, the usual result from an inference algorithm covers haplotype frequencies and possible haplotype phases of each genotype.

2.3. Haplotype Explanation Probability

In a genomic region with multiple heterozygous sites, multiple pairs of haplotypes can be inferred from a given genotype. The probability of a genotype to be phased into one specific pair of haplotypes would depend on the frequencies of haplotypes constituting the pairs (Becker et al., 2005). This probability is given by

$$w_{ij} = \frac{f_i f_j}{\sum_{(h_k, h_l) \in H} f_k f_l} \tag{2}$$

where w_{ij} is the probability for haplotype pair *ij*, f_i denotes the frequency of the *i*th haplotype, h_k is the *k*th haplotype and *H* represents the set of haplotype explanations which are compatible with the genotype of interest. For example, genotype `0110` can be phased into two haplotype pairs: `0000/0110` (h_1 / h_4) and `0010/0100` (h_2 / h_3). If the frequencies for haplotypes `0000`, `0010`, `0100` and `0110` are respectively 0.5, 0.2, 0.2 and 0.1, the probabilities for the pairs `0000/0110` and `0010/0100` are 0.556 and 0.444. Obviously, the probability of a genotype with all homozygous sites or single heterozygous site to be phased into a pair of haplotypes would be equal to one. In genetic interaction studies where the number of genes or unlinked regions is greater than one, the haplotype explanation probabilities from all regions can be combined together. An overall contribution by one sample to haplotype configuration $(h_j^1, h_j^2, ..., h_j^{n_u})$ in a study with n_u genes/unlinked regions is given by

$$c_{(h_j^1, h_j^2, ..., h_j^{n_u})} = 2\prod_{i=1}^{n_u} w_{jk}^i \frac{(1+\delta_{jk}^i)}{2} \tag{3}$$

where $c_{(h_j^1, h_j^2, ..., h_j^{n_u})}$ is the contribution value and δ is defined as $\delta_{jk} = 1$ for $j = k$ and $\delta_{jk} = 0$ for $j \neq k$. In the previous example where haplotypes from only one region are considered, $c_{h_1} = 0.556$, $c_{h_2} = 0.444$, $c_{h_3} = 0.444$ and $c_{h_4} = 0.556$. Notice that the sum of contribution values is equal to two; this reflects the fact that each genotype is made up from two haplotypes. Becker et al. (2005) use this contribution value in the construction of a contingency table where a χ^2 test statistic is subsequently calculated. With the use of a Monte Carlo simulation, an estimated *p*-value is then obtained for the test statistic. Similar to the model exploration strategy in MDR, the process of contingency table construction and *p*-value calculation can also be cycled through all or some of possible interaction models. The model with appropriate

candidate SNPs taken from possible SNPs is the one with minimum *p*-value and is said to be the best model for interaction explanation. This statistics-based procedure can be found as an integral part of the FAMHAP program (Becker and Knapp, 2004).

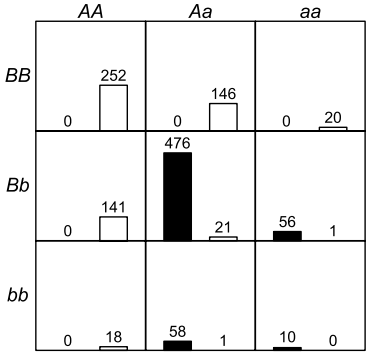

Figure 1. An MDR decision table which is constructed using 1,200 case-control samples. The genotype of each sample is determined from two SNPs. The table consists of nine cells where each cell represents a unique genotype. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AaBb*, *aaBb*, *Aabb* and *aabb* are labelled as predisposing genotypes while the cells with genotypes *AABB*, *AaBB*, *aaBB*, *AABb* and *AAbb* are labelled as protective genotypes.

3. VarHAP

VarHAP is proposed for case-control interaction studies. Similar to MDR, the technique is also a classifier-based technique. However, instead of using a genotype data analysis as a means to identify the best SNP combination, the decision table for classification is constructed from the haplotype contribution value described earlier. As a result, haplotypes with different lengths must be inferred during the search for the best model. The number of decision cells during the consideration on haplotypes constructed from a specific set of SNPs is governed by the total number of possible haplotype configurations as illustrated in Figure 2. In brief, VarHAP would maintain the ability to find the best SNP combination while also be able to identify possible disease-predisposing and protective haplotype configurations.

Since VarHAP is essentially a classification system, the principal criterion for choosing the optimal SNP combination model is still the prediction accuracy. However, with the use of haplotype contribution value as a means for decision rule construction, an additional model selection criterion that exploits the nature of haplotype can be formulated. This criterion can be referred to as haplotype propagation capability. Basically, if a haplotype constructed from a specific set of SNPs is related to disease susceptibility status, haplotypes constructed from a SNP set which is a superset of the previously specified SNPs should also predict the same relationship. This implies that predisposing and protective haplotypes in a low-order model must be able to propagate into haplotypes in high-order models. For example, consider a single-gene problem with four possible SNPs: `X1`, `X2`, `X3` and `X4`. If haplotypes in the model with SNPs `(X2, X4)` are related to disease susceptibility, haplotypes in the models with SNPs `(X1, X2, X4)`, `(X2, X3, X4)` and `(X1, X2, X3, X4)` should produce the same result. The haplotype propagation

capability, which is a dichotomous criterion, can be determined from the evidence that the sign test *p*-value and the prediction accuracy can be maintained throughout the process of model order increment. Again, in the situation where two or more models with different number of SNPs are equally good in terms of both prediction accuracy and haplotype propagation capability, the most parsimonious model will be the best model.

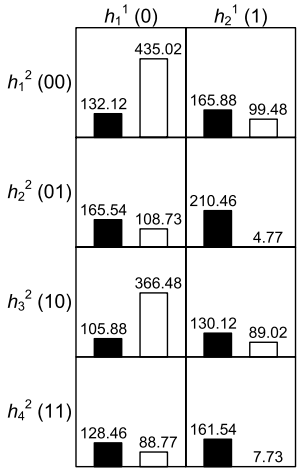

Figure 2. A VarHAP decision table which is constructed from 1,200 case-control samples. Haplotypes in the first gene are obtained from one SNP while haplotypes in the second gene are inferred from two SNPs. The table consists of eight cells where each cell represents a unique haplotype configuration. The left (black) bar in each cell represents the accumulative contribution from case samples while the right (white) bar represents the accumulative contribution from control samples. The cells with haplotype configurations (h_2^1, h_1^2), (h_1^1, h_2^2), (h_2^1, h_2^2), (h_2^1, h_3^2), (h_1^1, h_4^2) and (h_2^1, h_4^2) are labelled as predisposing haplotype configurations while the cells with haplotype configurations (h_1^1, h_1^2) and (h_1^1, h_3^2) are labelled as protective haplotype configurations.

4. Data Sets

The performance of the proposed VarHAP technique is evaluated through benchmark trials. 12 simulated data sets, which represent various gene-gene interaction phenomena including epistasis and heterogeneity, are considered (Knapp et al., 1994; Becker et al., 2005). Each data set contains 600 case samples and 600 control samples. Each sample consists of 10 total SNPs from two genes where five SNPs exist in each gene. All SNPs in control samples are in Hardy-Weinberg equilibrium (Hardy, 1908). Only one SNP from each gene is interacted with one another. The two-locus interaction models are illustrated in Table 1. The epistatic models Ep-1–Ep-6 and the heterogeneity models Het-1–Het-3 have been discussed by Neuman and Rice (1992), who also provide examples of diseases for which these models may be applicable. The heterogeneity models S-1 and S-2 and the epistatic model S-3 have been investigated by Schork et al. (1993). From Table 1, if the frequency of the disease allele at a locus is greater than 0.5, the major allele is the disease allele. Otherwise, the minor allele is the disease allele. These interaction models describe disease susceptibility status in terms of penetrance. Penetrance of a genotype with a specific number of disease alleles is the probability that a subject with this genotype has the disease. The test data sets are simulated by a genomeSIM package (Dudek et al., 2006) with the default setting. As a result, it is also possible to vary the LD pattern among SNPs in the same gene. This leads to two main case studies that need to be explored: strong LD and weak LD cases. In the strong LD case, the susceptibility-causative SNP in each gene and its two adjacent SNPs are in linkage disequilibrium where Lewontin's D' value (Lewontin, 1988) is in the range of 0.80–0.95. In contrast, the Lewontin's D' value for each pairwise LD measurement between susceptibility-causative SNP and its adjacent SNPs is in the range of 0.50–0.60 in the

weak LD case. In the strong LD case, an interaction detection technique should be able to identify both the actual two-locus model that directly leads to disease susceptibility and other alternative models which consist of SNPs in strong LD patterns. The ability to detect these other models is important. This is because it is not always straightforward to identify SNPs which are responsible for disease susceptibility in real case-control interaction studies. In contrast, an interaction detection technique should narrow the search to the original two-locus model in weak LD case since it is the only usable model.

*Table 1***.** Description of two-locus disease models. d_{ij} is the penetrance of a genotype carrying *i* disease alleles at locus 1 and *j* disease alleles at locus 2. p_1 is the frequency of the disease allele at locus 1 while p_2 is the frequency of the disease allele at locus 2. $\psi = 2\phi - \phi^2$.

Model	d_{22}	d_{21}	d_{20}	d_{12}	d_{11}	d_{10}	d_{02}	d_{01}	d_{00}	p_1	p_2	ϕ
Ep-1	ϕ	ϕ	0	ϕ	ϕ	0	0	0	0	0.210	0.210	0.707
Ep-2	ϕ	ϕ	0	0	0	0	0	0	0	0.600	0.199	0.778
Ep-3	ϕ	0	0	0	0	0	0	0	0	0.577	0.577	0.900
Ep-4	ϕ	ϕ	0	ϕ	0	0	ϕ	0	0	0.372	0.243	0.911
Ep-5	ϕ	ϕ	0	ϕ	0	0	0	0	0	0.349	0.349	0.799
Ep-6	0	ϕ	ϕ	ϕ	0	0	ϕ	0	0	0.190	0.190	1.000
Het-1	ψ	ψ	ϕ	ψ	ψ	ϕ	ϕ	ϕ	0	0.053	0.053	0.495
Het-2	ψ	ψ	ϕ	ϕ	ϕ	0	ϕ	ϕ	0	0.279	0.040	0.660
Het-3	ψ	ϕ	ϕ	ϕ	0	0	ϕ	0	0	0.194	0.194	1.000
S-1	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	0	0.052	0.052	0.522
S-2	1	1	1	ϕ	ϕ	0	ϕ	ϕ	0	0.228	0.045	0.574
S-3	1	1	ϕ	1	ϕ	0	ϕ	0	0	0.194	0.194	0.512

5. Results and Discussions

VarHAP is benchmarked against MDR and FAMHAP. Since the test data contains 10 SNPs, all three techniques have to explore $2^{10}-1=1,023$ possible SNP combination models. An initial investigation reveals that with the use of minimum *p*-value as the sole model selection criterion, FAMHAP reports a large number of models with the estimated *p*-value equals to zero. As a result, haplotype propagation capability is also implemented as an additional model selection criterion. Further, the parsimony criterion is also utilised when there is a tie between multiple models with different number of SNPs. The results from all three techniques in weak and strong LD case studies are summarised in Tables 2 and 3, respectively.

The prediction accuracy of MDR is higher than that of VarHAP in both case studies. This is because VarHAP uses contribution values which are obtained from inferred haplotypes instead of inferred diplotypes—pairs of haplotypes that together describe correct phases of given genotypes—to create decision rules. Consider a situation where disease susceptibility can be determined from a single SNP where the predisposing genotype is the homozygous variant. In other words, the disease susceptibility can be described by a recessive genetic model. MDR can easily classify the heterozygous and homozygous wide-type genotypes as protective genotypes. However, VarHAP would only correctly classify both homozygous genotypes since each genotype is made up from two copies of the same haplotype: two major alleles for the homozygous wide-type and two minor alleles for the homozygous variant. VarHAP would partially misclassify samples with heterozygous genotype. This is because VarHAP identifies the major allele as the protective allele and the minor allele as the predisposing allele. In order to increase the prediction accuracy of VarHAP, it may be necessary to construct decision tables from diplotype information

instead of haplotype contribution values. Nonetheless, this will also rapidly increase the dimensions of decision tables in VarHAP.

In the weak LD case study, both MDR and VarHAP are able to identify correct sets of SNPs that lead to disease susceptibility. On the other hand, FAMHAP reports both actual and alternative interaction models. This is undesirable since it would not be possible to further explain disease susceptibility from multiple candidate models in the absence of strong linkage disequilibrium among SNPs. In other words, FAMHAP is quite sensitive in this situation. Further analysis reveals that MDR is marginally better than VarHAP in two epistasis problems: Ep-2 and Ep-5. MDR correctly identifies models which contain two SNPs while the models located by VarHAP contain a few extra SNPs. Nonetheless, these two models identified by VarHAP are still useful to susceptibility explanation.

All three techniques are able to locate correct interaction models in the strong LD case study. However, only FAMHAP and VarHAP are capable of identifying alternative models. Since MDR suggests one candidate model for each fixed-number SNP set, it would not be possible for MDR to produce any alternative models. Recall that these alternative models are equally important since SNPs in the principal two-locus interaction model and SNPs from an alternative model are in strong linkage disequilibrium. This implies that disease susceptibility can be explained using either the original interaction model or the alternative models. This disadvantage in MDR can be overcome if the cross-validation consistency criterion can be replaced by other decision criteria. In this case study, FAMHAP is marginally better than VarHAP in terms of alternative model identification in three epistasis and heterogeneity problems: Ep-3, Ep-6 and Het-3. This means that FAMHAP is at its best when SNPs are in strong linkage disequilibrium. Nonetheless, the overall results from both case

studies suggest that VarHAP is the best technique. This is concluded from the fact that VarHAP does not report ambiguous results in weak LD case study while is also capable of producing alternative models in strong LD case study. This is crucial because it is impossible to know beforehand whether susceptibility-causative SNPs are in weak or strong linkage disequilibrium with other SNPs in real case-control interaction studies. In other words, a technique that performs satisfactorily in both weak and strong LD cases would have an advantage over a technique that functions well in only one scenario.

Table 2. MDR, VarHAP and FAMHAP results from the weak LD case study. 10-fold cross-validation is used in MDR and VarHAP. The prediction accuracy is obtained for the identified principal interaction model. Estimated *p*-values in FAMHAP results are equal to zero while sign test *p*-values in MDR and VarHAP results are less than 0.001 in all two-locus problems. The technique is said to be able to identify the correct gene-gene interaction model if the reported principal model contains both SNPs which are directly participated in the interaction model. Alternative models are models which contain at least two SNPs where each SNP must be either a SNP from the two-locus model or a SNP which is in linkage disequilibrium with one of the SNPs from the model. The number in each bracket denotes the order of the identified model (the number of SNPs in the model).

Two-Locus Model	MDR Prediction Accuracy (%)	VarHAP Prediction Accuracy (%)	Correct Model Identification Technique	Alternative Model Identification Technique
Ep-1	98.00	73.92	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-2	98.58	78.39	MDR(2), VarHAP(4), FAMHAP(2)	FAMHAP(2)
Ep-3	99.50	87.50	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-4	99.25	78.96	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-5	98.42	75.19	MDR(2), VarHAP(3), FAMHAP(2)	FAMHAP(2)
Ep-6	100.00	85.10	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Het-1	93.75	73.29	MDR(2), VarHAP(2), FAMHAP(2)	
Het-2	97.33	78.40	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Het-3	100.00	84.40	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
S-1	94.00	72.98	MDR(2), VarHAP(2), FAMHAP(2)	
S-2	97.58	79.81	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
S-3	96.75	79.15	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)

Table 3. MDR, VarHAP and FAMHAP results from the strong LD case study. The explanation for how the results are obtained and displayed is the same as that given in Table 2.

Two-Locus Model	MDR Prediction Accuracy (%)	VarHAP Prediction Accuracy (%)	Correct Model Identification Technique	Alternative Model Identification Technique
Ep-1	98.00	73.92	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
Ep-2	98.58	77.02	MDR(2), VarHAP(4), FAMHAP(2)	VarHAP(4), FAMHAP(2)
Ep-3	99.50	87.50	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-4	99.25	78.96	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
Ep-5	98.42	75.87	MDR(2), VarHAP(3), FAMHAP(2)	VarHAP(3), FAMHAP(2)
Ep-6	100.00	85.10	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Het-1	93.75	75.41	MDR(2), VarHAP(3), FAMHAP(2)	VarHAP(3), FAMHAP(2)
Het-2	97.33	78.40	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
Het-3	100.00	84.40	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
S-1	94.00	72.98	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
S-2	97.58	79.81	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
S-3	96.75	79.15	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)

6. Conclusions

In this report, a non-parametric pattern recognition/classification technique for case-control gene-gene interaction studies is presented. Instead of using direct genotype inputs in classification, inferred haplotypes, which are obtained through an expectation-maximisation algorithm (Excoffier and Slatkin, 1995), are used as inputs. Each case/control sample contributes values derived from inferred haplotypes to decision tables which are constructed and tested for all possible gene-gene interaction models. The technique primarily uses prediction accuracy obtained from *k*-fold cross-validation as a means for identifying candidate SNPs which are responsible for disease susceptibility. The technique also employs haplotype propagation capability as an additional criterion. If the selection procedure ends in a tie between two or more models with different number of SNPs, the most parsimonious model is then reported as the interaction model. Since haplotypes with different length must be constructed during model identification, the proposed technique can be referred to as a variable-length haplotype construction for gene-gene interaction (VarHAP) technique. VarHAP has been benchmarked against two interaction model detection programs namely MDR (Hahn et al., 2003) and FAMHAP (Becker and Knapp, 2004; Becker et al., 2005) in 12 two-locus epistasis and heterogeneity problems (Knapp et al., 1994; Becker et al., 2005). The results reveal that FAMHAP reports multiple ambiguous models in the presence of weak linkage disequilibrium among input SNPs while MDR is not suitable for alternative interaction model identification when input SNPs are in strong linkage disequilibrium. In contrast, VarHAP emerges as the most suitable technique in both situations involving weak and strong linkage disequilibrium. Suggestions for further improvement of MDR and VarHAP are also included.

Supplementary Information

VarHAP, which is implemented in Java, and the simulated data sets used in the report are available upon request (e-mail: nchl@kmutnb.ac.th). In addition to the use of the genomeSIM package (Dudek et al., 2006), the data sets can also be generated by a SNaP package (Nothnagel, 2002). Readers might also be interested in applying the techniques discussed in this report to examples of case-control data sets, which are publicly available from the Wellcome Trust Case Control Consortium (The Wellcome Trust Case Control Consortium, 2007).

Output from the Project

The research results have been published in two international journal articles; details of these articles follow.

1. Thongngarm, T., Jameekornrak, A., Limwongse, C., Sangasapaviliya, A., Jirapongsananuruk, O., Assawamakin, A., Chaiyaratana, N., Luangwedchakarn, V. and Thongnoppakhun, W. (2008). Association between *ADAM33* polymorphisms and asthma in a Thai population. *Asian Pacific Journal of Allergy and Immunology*, *26*, 205−211 (2007 Journal Impact Factor = 0.567).
2. Assawamakin, A., Chaiyaratana, N., Limwongse, C., Sinsomros, S., Yenchitsomanus, P.-T. and Youngkong, P. (2009). Variable-length haplotype construction for gene-gene interaction studies. *IEEE Engineering in Medicine and Biology Magazine*, *28*, in press (2007 Journal Impact Factor = 1.066).

Appendix

Publication of the Research Results

A.1. Asian Pacific Journal of Allergy and Immunology

Thongngarm, T., Jameekornrak, A., Limwongse, C., Sangasapaviliya, A., Jirapongsananuruk, O., Assawamakin, A., Chaiyaratana, N., Luangwedchakarn, V. and Thongnoppakhun, W. (2008). Association between *ADAM33* polymorphisms and asthma in a Thai population. *Asian Pacific Journal of Allergy and Immunology*, *26*, 205−211 (2007 Journal Impact Factor = 0.567).

ASIAN PACIFIC JOURNAL OF ALLERGY AND IMMUNOLOGY (2008) 26: 205-211

Association between *ADAM33* Polymorphisms and Asthma in a Thai Population

Torpong Thongngarm[1], Aree Jameekornrak[1], Chanin Limwongse[2,5], Atik Sangasapaviliya[3], Orathai Jirapongsananuruk[4], Anunchai Assawamakin[5], Nachol Chaiyaratana[5,6], Voravich Luangwedchakarn[7] and Wanna Thongnoppakhun[5]

SUMMARY *ADAM33 (*A Disintegrin And Metalloprotease 33) is an asthma susceptibility gene found across several human populations**.** However, no information on *ADAM33* exists for Thai population. The objective of this study was to determine the association, if any, between *ADAM33* polymorphisms and asthma in Thai subjects. Genotyping revealed 8 single nucleotide polymorphisms (SNPs) within the 3′ region of the *ADAM33* gene among 200 asthmatics and 100 control subjects. Asthmatic subjects were further sub-categorized into high and low severity groups. Multiple genetic model statistic tests for single-marker and haplotype association were carried out. Differences in allele frequencies at the SNPs rs528557/S2, rs598418 and rs44707/ST+4 in asthmatics were statistically significant compared to controls. The SNP rs528557/S2 could also be linked to the low severity group and the SNPs rs598418 and rs44707/ST+4 with the high severity group. Two-SNP haplotype analysis at the SNPs rs528557/S2 and rs598418 revealed a significant association with asthma. This study in a Thai population confirmed a positive association between *ADAM33* polymorphisms and asthma susceptibility.

Asthma is a common chronic respiratory disease in which chronic inflammation leads to an airflow obstruction and bronchial hyperresponsiveness (BHR) and results in irreversible structural changes in airway remodeling.[1] The cause of this disease is believed to be a complex combination of multiple genetic and environmental factors which lead to heterogeneous phenotypes such as variable degrees of atopic involvement and severity.[2] Early family and twin studies support the role of genetics in the development of the disease.[3,4] A large number of association and linkage studies have identified over 100 genes related to asthma. However, less than a dozen of these are associated with asthma, according to a large number of independent reports.[5] This group of highly promising candidates, discovered through genome-wide linkage analysis, consists of known pathogenesis-related genes and also novel genes of unprecedented linkage to asthma.

The first positionally-cloned asthma susceptibility gene is a member of the *ADAM* (A Disintegrin And Metalloprotease) gene family, *ADAM33*, which is located on human chromosome 20p.[6] Ge-

From the [1]Division of Allergy and Clinical Immunology, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, [2]Division of Medical Genetics, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, [3]Allergy and Immunology Unit, Department of Medicine, Pramongkutklao Hospital, Bangkok, [4]Division of Allergy and Clinical Immunology, Department of Pediatrics, Faculty of Medicine Siriraj Hospital, Mahidol University, [5]Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, [6]Department of Electrical Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, [7]Department of Immunology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.
Correspondence: Torpong Thongngarm
E-mail: sittn@mahidol.ac.th

netic linkage analysis and association studies of families with asthma across diverse ethnic backgrounds support a relationship between *ADAM33* polymorphisms and asthma phenotypes and airway hyperreactivity (AHR).[7-10] The mouse orthologue of *ADAM33* also lies in the region linked to AHR.[11] The precise mechanisms behind polymorphic variations in *ADAM33* or their encoded proteins in contributing to asthma development remain under intense investigation. Selective expression of the *ADAM33* mRNA and its protein in adult bronchial smooth muscle and human embryonic bronchi and surrounding mesenchyme indicate specific roles of the gene in the observed phenotypes of AHR and airway remodeling.[12-14]

Despite the above notion, negative associations between *ADAM33* polymorphisms with asthma have been reported from ethnically different populations.[15-17] Among various factors, a heterogeneity of allelic frequencies at certain loci among samples from these groups may be yielding discordant results. There is an increased prevalence of asthma in Thailand which may be following similar trends observed in many Western and Asian countries. However, the differences in the genetic background of the Thai population may account for different asthma susceptibility genes.[18] Therefore, we sought to determine *ADAM33* polymorphisms in Thai asthmatics and compared them with non-asthmatic counterparts.

MATERIALS AND METHODS

Subjects

One hundred control subjects and 200 asthmatics of Thai nationality were recruited from pediatric and adult allergy and immunology clinics from two tertiary care centers in Bangkok, Thailand. Asthma was diagnosed based on symptoms and on spirometry assessments using the criteria outlined by the American Thoracic Society. Degrees of asthma severity were determined according to guidelines provided by the Global Initiative for Asthma.[19] Control subjects were asymptomatic for asthma and were devoid of atopic or pulmonary diseases. Pregnant or lactating female subjects were excluded. The study was in accordance with the Helsinki Declaration and approved by the local institutional review boards, and all subjects provided informed, written consent.

Polymorphism genotyping

ADAM33 polymorphisms were genotyped in the 3′ region, from exon 19 to exon 22, by direct sequencing of genomic DNA extracted from peripheral blood leukocytes. The PCR primers designed by the authors are shown in Table 1. They were used to amplify gene segments that spanned eight previously reported SNP locations, which are also known to be associated with the asthma phenotype across ethnically diverse populations.[7] These SNPS are rs3918396/S1, rs528557/S2, rs44707/ST+4, rs574174/ST+7, rs2280091/T1, rs2280090/T2, rs543749/V-1 and rs2787094/V4. PCR was performed in a 25-µl reaction volume containing 1x buffer, 1.0 mM $MgCl_2$, 200 µM dNTPs, 1x Q-solution, 1.0 µM of each primer, 1.0 U *Taq* polymerase (Qiagen, Hilden, Germany) and 50 ng of genomic DNA. Reactions consisted of 30 cycles of denaturation at 94ºC for 30 seconds, annealing at the optimal melting temperature for 30 seconds, and extension at 72ºC for 1 minute on a PTC-100 Programmable Thermal Controller (MJ Research, Waltham, USA). PCR products were

Table 1 Oligonucleotide primers used for resequencing the 3' region of *ADAM33*

Position	Forward primer 5'→3'	Reverse primer 5'→3'
Exon 19	5'-TGACTGCCTGCCACAGCCAC-3'	5'-TCTGAGGAGGGGAACCGCAG-3'
Intron 19-1	5'-GCAGTGGGTAGGCTCCGAGC-3'	5'-AGAGTGCCTGCCCTGCCTAG-3'
Intron 19-2	5'-GCGGAGTGGGGAGTCACATAATAC-3'	5'-GGCTGGCACCTCCTCTCTCTAG-3'
Exon 20, 21	5'-AGGTTCTTTGGAAGCTGAGCG-3'	5'-ACTGAGGGGTGGGAGAGGTG-3'
Intron 21	5'-GGCAGGGACCTGGATTCAAAG-3'	5'-CACACCAGACTCCCAGGACAGAG-3'
Exon 22	5'-GTCCCAGAAGCAAAGGTCACAC-3'	5'-TGCGGTGTCTTGCTGTGTTG-3'

purified by the PCR purification kit (Qiagen, Hilden, Germany) and were used as DNA template for cycle sequencing. Direct DNA sequencing was performed using a DYEnamic ET Terminator cycle sequencing kit (Amersham Bioscience, UK). Sequencing reactions were prepared in 10-µl volumes containing 1.0 µM primer, 5 pmol DYE-ET and 30 ng of DNA template and subjected to 25 cycles of denaturation at 95ºC for 20 seconds, annealing at 50ºC for 15 seconds, and extension at 60ºC for 1 minute on a PTC-100 Programmable Thermal Controller.

Statistical analysis

The distribution of allele frequencies was tested for conformity to the Hardy-Weinberg equilibrium by χ^2 test. The degree of linkage disequilibrium between loci were measured using Lewontin's D'.[20] Each SNP was analyzed by comparing differences in genotype frequencies between those in the asthmatic and control groups stratified by population, and χ^2 tests for association were performed.[21] In order to analyze the association between the severity of asthma and *ADAM33* polymorphisms, we subcategorized asthmatic subjects into two groups, termed the high and low severity groups. The high severity group included subjects suffering from moderate to severe persistent asthma, whereas the low severity group consisted of subjects with intermittent and mild persistent asthma. The degree of asthma severity was classified as per the guidelines provided by the Global Initiative for Asthma.

Haplotype association analysis

Haplotype association analysis was carried out in two steps. An optimal SNP set for association was first identified by single marker analysis. Then, in the second step, genetic models for haplotype association based on the optimal SNP set were identified using HAPSTAT.[22-24] Likelihood ratio tests were performed for each genetic model.

RESULTS

One hundred control subjects (46 men and 54 women, mean age 26 years) and 200 individuals suffering from asthma (116 men and 84 women, mean age 29.86 years) were recruited and genotyped for *ADAM33* polymorphisms. The control group had significantly higher lung function parameters than the asthmatic group. Asthmatic subjects showed a mean FEV_1 (forced expiratory volume in 1 second) of 1.83 l (66.34% of predicted) and a mean FVC (forced vital capacity) of 2.89 l (80.14% of predicted). The mean duration of asthma was 10.68 years. Table 2 shows that subjects in the high severity group were significantly older and suffered from asthma for longer durations than those in the low severity group. The mean FEV_1 was significantly lower in the high severity group as compared to those of the low severity group.

Genotyping analyses revealed eight SNPs in the 3′ region of *ADAM33*, being distributed in Hardy-Weinberg equilibrium. Only one SNP, rs528557/S2, was located on exon 19, while the other seven SNPs (rs2853209, rs598418, rs44707/ST+4, rs597980, rs11905233, rs2787094/V4 and rs3746631) were located on intron 19 and 3' UTR. Genotype descriptions for each SNP are outlined in Table 3.

Table 2 Clinical characteristics of high and low severity groups

Characteristics*	Low severity group (n = 95)	High severity group (n = 105)	p-value
Age (years)	22.94 (17.99)	35.78 (17.06)	< 0.0001
Gender (M/F ratio)	1.21	1.56	NS
Duration of asthma (years)	8.91 (6.91)	12.20 (8.23)	0.002
FVC (% of predicted)	81.33 (21.15)	79.19 (16.57)	NS
FEV_1 (% of predicted)	72.27 (16.27)	61.26 (15.83)	< 0.0001

*For age, duration of asthma, FVC and FEV_1 characteristics, the displayed values are means while the numbers in brackets are SDs. For gender characteristics, the displayed values are the ratios between males and females. p-values are obtained from t-tests except for the gender characteristic where the p-value is calculated using a Fisher's exact test. FVC, forced vital capacity; FEV_1, forced expiratory volume in 1 second; NS, non-significant.

Single SNP analysis

Table 4 indicates that there are significant differences in the genotypes observed in asthmatics and in control groups for the SNPs rs528557/S2 (p = 0.012), rs598418 (dominant model: p = 0.017, multiplicative model: p = 0.049) and rs44707/ST+4 (p = 0.049). The asthmatic patients were divided into high and low severity subgroups based on clinical characteristics, as previously described. There was a significant difference in allele frequencies between the low severity and control groups for the SNP rs528557/S2 (p = 0.012). Differences in allelic distribution were also significant between the high severity and control groups for the SNPs rs598418 (p = 0.017) and rs44707/ST+4 (p = 0.045). No statistically significant differences in allele and genotype frequencies between high and the low severity groups were observed.

Haplotype association analysis

The frequency of the haplotype CT at the SNPs rs528557/S2 and rs598418 of asthmatic patients was significantly higher than that of controls (p = 0.036), as shown in Table 5A. Fig. 1 shows that these two SNPs are in strong LD (D' = 0.93995). At the SNPs rs598418 and rs44707/ST+4, a significantly higher frequency of the haplotype CC in the control group when compared to the high severity group was observed (p = 0.046) whereas the haplotype TA was demonstrated to be an at-risk haplotype for the high severity group (p = 0.046), as shown in Table 5B. The SNPs rs598418 and rs44707/ST+4 are also in strong LD (D' = 0.98251) (Fig. 1). No significant differences in the haplotype frequencies between the low severity group and the control group were observed.

Table 3 Genotype description for *ADAM33* SNPs in a Thai population

SNP ID	SNP type	Genotype, case (n = 200)/control (n = 100)		
		Homozygous wild-type	Heterozygous	Homozygous variant
rs528557/S2	Exon	GG (114/72)	GC (77/21)	CC (9/7)
rs2853209	Intron	AA (67/42)	AT (106/45)	TT (27/13)
rs598418	Intron	CC (46/36)	CT (102/42)	TT (52/22)
rs44707/ST+4	Intron	CC (63/43)	CA (103/42)	AA (34/15)
rs597980	Intron	CC (91/53)	CT (97/38)	TT (12/9)
rs11905233	3′ UTR	GG (193/100)	GA (6/0)	AA (1/0)
rs2787094/V4	3′ UTR	CC (87/40)	CG (94/47)	GG (19/13)
rs3746631	3′ UTR	AA (163/88)	AG (37/12)	GG (0/0)

Table 4 Association of SNPs in *ADAM33*

SNP ID	Genetic model	Genotype	Allele	Case-control	High severity-control	Low severity-control
				p-value	*p*-value*	*p*-value*
rs528557/S2	Dominant	GC + CC GG		0.012	NS	0.012
rs598418	Dominant	CT + TT CC		0.017	0.017	NS
rs598418	Multiplicative		T C	0.049	NS	NS
rs44707/ST+4	Dominant	CA+AA CC		0.049	0.045	NS

*NS, non-significant.

DISCUSSION

We successfully amplified gene segments covering eight SNPs previously reported in other populations.[7] We found only three SNPs in the studied samples that matched those previously reported, i.e. rs528557/S2, rs44707/ST+4 and rs2787094/V4. The SNP rs528557/S2 exhibited significant association with asthma in a dominant model (p = 0.012) whereas the SNP rs44707/ST+4 was marginally associated to the disease according to a dominant model (p = 0.049). The SNP rs528557/S2 has been linked to asthma within UK,[6] German,[8] African-American, US Caucasian and US Hispanic[7] populations. The SNP rs44707/ST+4 was also linked to asthma in the UK and in combined US and UK populations.[6] The SNP rs598418 was associated with asthma in both multiplicative and dominant models (p = 0.049 and 0.017, respectively). Other SNPs, including rs2853209, rs597980, rs11905233 and rs3746631, were not significantly associated with the prevalence of asthma in this Thai population. However, there is a report on the SNP rs597980 as having a positive association with asthma and BHR in a German population.[8] Association between *ADAM33* SNPs and asthma has also been demonstrated in other Asian populations, such as Korean[9] and Japanese.[25,26] A meta-analysis study involving eight populations confirmed the positive association be-

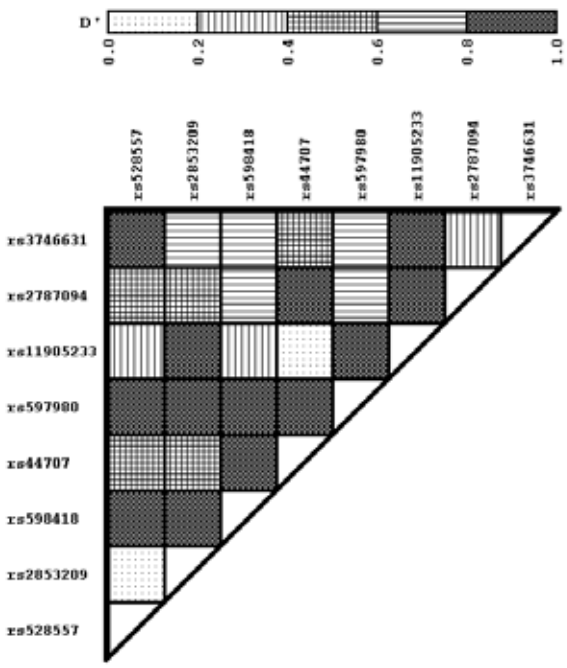

Fig. 1 Pair-wise linkage disequilibrium analysis in the 3' region of *ADAM33* estimated in a Thai population.

Table 5 Haplotype analysis of *ADAM33* (two-SNP haplotype)

A. Case *vs.* control (rs528557/S2 and rs598418)

Haplotype	Haplotype frequency			*p*-value*
	Control (n = 100)	**Case** (n = 200)	**Combined** (n = 300)	
GC	0.56	0.48	0.50	NS
GT	0.26	0.28	0.28	NS
CC	0.01	0.01	0.01	NS
CT	0.17	0.23	0.21	0.036 (dominant)

*NS, non-significant.

B. High severity *vs.* control (rs598418 and rs44707/ST+4)

Haplotype	Haplotype frequency			*p*-value*
	Control (n = 100)	**High severity** (n = 105)	**Combined** (n = 205)	
CC	0.57	0.49	0.53	0.046 (recessive)
CA	0.00	0.00	0.00	NS
TC	0.07	0.09	0.08	NS
TA	0.36	0.42	0.39	0.046 (joint dominant)

*NS, non-significant.

tween *ADAM33* polymorphisms and asthma susceptibility.[10] In contrast, a lack of association between *ADAM33* polymorphisms and asthma have been reported from ethnically different populations.[15-17] This may be influenced by sample sizes, heterogeneity of populations, and different environmental exposures. It is interesting to note that only the SNP rs528557/S2 is located on exon 19 while the other seven SNPs are located on intron 19 and 3' UTR. A previous study has detailed the mechanisms behind the regulation of the 3' UTR of *ADAM33* involving mRNA localization and processing, as well as protein maturation.[27]

Our study also demonstrated a significant association between *ADAM33* SNPs and asthma severity. When compared to the control group, the SNP rs528557/S2 was associated with the low severity group (p = 0.012) whereas the SNPs rs598418 and rs44707/ST+4 were associated with the high severity group (p = 0.017 and 0.045, respectively). Subjects in the high severity group were significantly older and suffered longer durations with asthma than those in the low severity group. This finding was in accordance with previous work revealing that lung function in asthmatic patients may decline over time.[28] Haplotypes can reveal information about hitherto unobserved predisposing variants in the region.[29] According to a dominant model, the haplotype CT at the SNPs rs528557/S2 and rs598418 was found to be a disease-predisposing variant in the Thai population. Upon comparison of the high severity and the control group, significant haplotypes were found at the SNPs rs598418 and rs44707/ST+4 (CC, p = 0.046; TA, p = 0.046). This result supports a previous study demonstrating significant global haplotypic association with asthma and asthma severity.[30] Notably, the χ^2 tests (with the application of Bonferroni correction to the single-marker analysis) failed to show any association of *ADAM33* and asthma in this study. Nevertheless, all reported p-values in our haplotype analysis were corrected by HAPSTAT[22-24] and the results confirmed statistical significance. This suggests that both single-marker and haplotype analyses should be performed in association studies, since haplotype-based methods incorporate linkage disequilibrium data from multiple markers and are more powerful for gene mapping than single SNP-based methods.

We did not find significant differences in the distribution of the SNPs and haplotypes between the high and low severity groups, possibly due to the rather small sample sizes. However, a cohort of 200 asthma patients monitored over a 20 year period showed that *ADAM33* SNPs could be linked to a decline in FEV_1, suggesting that the gene may not only affect asthma prevalence but also its severity and disease progression.[31] A potential role of *ADAM33* in asthma pathogenesis may involve the mRNA and protein expression in mesenchymal cells, airway smooth muscle cells and fibroblasts.[12,27] Previous studies have shown a link between *ADAM33* regulation and the expression of TGF-β, the important cytokine produced by bronchial epithelial cells for repair processes.[32] ADAM33 protein levels were significantly elevated in patients with asthma[14] and, in turn, would also involve BHR and airway remodeling.

It is possible that the Thai subjects in the present study are of other ethnic origins. However, we believe that our findings were not affected by population stratification, for two reasons. Firstly, all eight SNPs in control samples were in Hardy-Weinberg equilibrium. Secondly, all SNPs were in strong linkage disequilibrium. Previous studies in other ethnic groups including German,[8] Korean,[9] and Japanese[25] have also indicated that SNPs within the 3′ region of *ADAM33* gene are in strong linkage disequilibrium. The evidence that recombination does not occur frequently in this genomic region among various ethnic groups implies that if population stratification occurs, deviation from Hardy-Weinberg equilibrium would be easily detectable.

In summary, we have identified a positive relationship between *ADAM33* polymorphisms and asthma in a Thai population. The results are supportive of previous studies involving other populations worldwide. The precise roles of *ADAM33* gene in asthma pathogenesis, however, remain undefined and require further investigation.

ACKNOWLEDGEMENTS

This study was supported by a research grant from the Faculty of Medicine Siriraj Hospital, Mahidol University (to TT) and Mahidol Research Grant (to CL). AA was supported by the Thailand Research

Fund (TRF) through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/4.I.MU.45/C.1) and the National Center for Genetic Engineering and Biotechnology (BIOTEC), the National Science and Technology Development Agency (NSTDA). NC was supported by the Thailand Research Fund through the Research Career Development Grant (Grant No. RSA5180006).

REFERENCES

1. Cockcroft DW, Davis BE. Mechanisms of airway hyperresponsiveness. J Allergy Clin Immunol 2006; 118: 551-9.
2. Holgate ST, Davies DE, Powell RM, Howarth PH, Haitchi HM, Holloway JW. Local genetic and environmental factors in asthma disease pathogenesis: chronicity and persistence mechanisms. Eur Respir J 2007; 29: 793-803.
3. Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD. Genetics of asthma and hay fever in Australian twins. Am Rev Respir Dis 1990; 142(6 Pt 1): 1351-8.
4. Skadhauge LR, Christensen K, Kyvik KO, Sigsgaard T. Genetic and environmental influence on asthma: a population-based study of 11,688 Danish twin pairs. Eur Respir J 1999; 13: 8-14.
5. Ober C, Hoffjan S. Asthma genetics 2006: the long and winding road to gene discovery. Genes Immun 2006; 7: 95-100.
6. Van Eerdewegh P, Little RD, Dupuis J, *et al.* Association of the *ADAM33* gene with asthma and bronchial hyperresponsiveness. Nature 2002; 418: 426-30.
7. Howard TD, Postma DS, Jongepier H, *et al.* Association of a disintegrin and metalloprotease 33 (*ADAM33*) gene with asthma in ethnically diverse populations. J Allergy Clin Immunol 2003; 112: 717-22.
8. Werner M, Herbon N, Gohlke H, *et al.* Asthma is associated with single-nucleotide polymorphisms in *ADAM33*. Clin Exp Allergy 2004; 34: 26-31.
9. Lee JH, Park HS, Park SW, *et al. ADAM33* polymorphism: association with bronchial hyper-responsiveness in Korean asthmatics. Clin Exp Allergy 2004; 34: 860-5.
10. Blakey J, Halapi E, Bjornsdottir US, *et al.* Contribution of *ADAM33* polymorphisms to the population risk of asthma. Thorax 2005; 60: 274-6.
11. De Sanctis GT, Merchant M, Beier DR, *et al*. Quantitative locus analysis of airway hyperresponsiveness in A/J and C57BL/6J mice. Nat Genet 1995; 11: 150-4.
12. Haitchi HM, Powell RM, Shaw TJ, *et al. ADAM33* expression in asthmatic airways and human embryonic lungs. Am J Respir Crit Care Med 2005; 171: 958-65.
13. Simpson A, Maniatis N, Jury F, *et al.* Polymorphisms in a disintegrin and metalloprotease 33 (*ADAM33*) predict impaired early-life lung function. Am J Respir Crit Care Med 2005; 172: 55-60.
14. Lee JY, Park SW, Chang HK, *et al.* A disintegrin and metalloproteinase 33 protein in patients with asthma: Relevance to airflow limitation. Am J Respir Crit Care Med 2006; 173: 729-35.
15. Lind DL, Choudhry S, Ung N, *et al*. *ADAM33* is not associated with asthma in Puerto Rican or Mexican populations. Am J Respir Crit Care Med 2003; 168: 1312-6.
16. Raby BA, Silverman EK, Kwiatkowski DJ, Lange C, Lazarus R, Weiss ST. *ADAM33* polymorphisms and phenotype associations in childhood asthma. J Allergy Clin Immunol 2004; 113: 1071-8.
17. Schedel M, Depner M, Schoen C, *et al.* The role of polymorphisms in *ADAM33*, a disintegrin and metalloprotease 33, in childhood asthma and lung function in two German populations. Respir Res 2006; 7: 91.
18. Trakultivakorn M, Sangsupawanich P, Vichyanond P. Time trends of the prevalence of asthma, rhinitis and eczema in Thai children-ISAAC (International Study of Asthma and Allergies in Childhood) Phase Three. J Asthma 2007; 44: 609-11.
19. National Institutes of Health and National Heart Lung and Blood Institute. Global strategy for asthma management and prevention guidelines. Global Initiative for Asthma 2006.
20. Hedrick PW. Gametic disequilibrium measures: proceed with caution. Genetics 1987; 117: 331-41.
21. Lewis CM. Genetic association studies: design, analysis and interpretation. Brief Bioinform 2002; 3: 146-53.
22. Lin DY, Zeng D, Millikan R. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 2005; 29: 299-312.
23. Lin DY, Zeng, D. Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc 2006; 101: 89-118.
24. Zeng D, Lin DY, Avery CL, North KE, Bray MS. Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. Biostatistics 2006; 7: 486-502.
25. Noguchi E, Ohtsuki Y, Tokunaga K, *et al. ADAM33* polymorphisms are associated with asthma susceptibility in a Japanese population. Clin Exp Allergy 2006; 36: 602-8.
26. Hirota T, Hasegawa K, Obara K, *et al.* Association between *ADAM33* polymorphisms and adult asthma in the Japanese population. Clin Exp Allergy 2006; 36: 884-91.
27. Umland SP, Garlisi CG, Shah H, *et al.* Human *ADAM33* messenger RNA expression profile and post-transcriptional regulation. Am J Respir Cell Mol Biol 2003; 29: 571-82.
28. Lange P, Parner J, Vestbo J, Schnohr P, Jensen G. A 15-year follow-up study of ventilatory function in adults with asthma. N Engl J Med 1998; 339: 1194-200.
29. Cordell HJ. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. Genet Epidemiol 2006; 30: 259-75.
30. Kedda MA, Duffy DL, Bradley B, O'Hehir RE, Thompson PJ. *ADAM33* haplotypes are associated with asthma in a large Australian population. Eur J Hum Genet 2006; 14: 1027-36.
31. Jongepier H, Boezen HM, Dijkstra A, *et al*. Polymorphisms of the *ADAM33* gene are associated with accelerated lung function decline in asthma. Clin Exp Allergy 2004; 34: 757-60.
32. Davies DE, Wicks J, Powell RM, Puddicombe SM, Holgate ST. Airway remodeling in asthma: new insights. J Allergy Clin Immunol 2003; 111: 215-25.

A.2. IEEE Engineering in Medicine and Biology Magazine

Assawamakin, A., Chaiyaratana, N., Limwongse, C., Sinsomros, S., Yenchitsomanus, P.-T. and Youngkong, P. (2009). Variable-length haplotype construction for gene-gene interaction studies. *IEEE Engineering in Medicine and Biology Magazine*, *28*, in press (2007 Journal Impact Factor = 1.066).

- Unconditional Acceptance -

To: nchl@kmitnb.ac.th,n.chaiyaratana@gmail.com
From: jenderle@bme.uconn.edu
Date: Nov 19, 2007
CC: ASJagath@ntu.edu.sg

Subject: RE: Manuscript No. EMB_ Mag-00054-2007.R1

Dear Dr. Chaiyaratana

Congratulations. Your manuscript, Variable-Length Haplotype Construction for Gene-Gene Interaction Studies, has been reviewed by expert referees, and the review suggests ACCEPTANCE. I have enclosed their comments so that you may fine tune the manuscript even further. If there is no comments, you can consider the paper accepted as is.

The next step is to prepare your paper for final submission. I would like you to send the following to the guest/associate editor as soon as possible but no longer than 30 days from today,

- One paper copy of the final manuscript, including clean hardcopy of each of the figures and Author Checklist (found at http://EMB-Magazine.bme.uconn.edu under author guidelines).

- An electronic copy of the paper (Microsoft Word file, PC Version) without figures, photographs or graphs (just text). An electronic PDF copy of the paper in a file with text and inserted figures with captions. Provide figures as separate files labeled Fig1, Fig2, etc., with format identified in the extension. Submit the files on a CD or one (or more) diskettes (floppy or Zip). TeX/LaTeX is not acceptable.

-1 author photo/ biography kit for each author of the paper

- The copyright form is to be submitted electronically via manuscript central.

Further information about getting your paper ready for publication is available at the website http://EMB_Magazine.bme.uconn.edu under author guidelines. If you have any questions please write to me at jenderle@bme.uconn.edu.

The EMB Magazine encourages authors to pay the voluntary page charge of $110 per page, which entitles the author to 100 reprints. Also color illustrations cost $150 per figure. By

submitting a figure with color, the author agrees to pay $150 for each color figure. There is a MANDATORY PAGE CHARGE at $250 per page in excess of the first five published pages. Payment for the mandatory page charge is not negotiable or voluntary.

Once again congratulations. I appreciate your selection of the EMB Magazine to publish your work, and I look forward to receiving the final version of your manuscript.

Sincerely yours,

John D. Enderle
Editor-in-Chief, EMB Magazine

University of Connecticut
260 Glenbrook Road, Room 217
Storrs, CT 06269-2247
Phone: (860) 486-5521

Associate Editor comments
==========
The figure and table captions could be shortened without losing the independence from the text.

Close Window

Variable-Length Haplotype Construction for Gene-Gene Interaction Studies

Journal:	*Engineering in Medicine and Biology Magazine*
Manuscript ID:	EMB_ Mag-00054-2007.R1
Manuscript Type:	By Invitation Only: Special Issue: Pattern extraction from bioinformatics data. Guest Editor: Jagath C Rajapakse
Date Submitted by the Author:	02-Nov-2007
Complete List of Authors:	Assawamakin, Anunchai; Mahidol University, Department of Research and Development Chaiyaratana, Nachol; King Mongkut's Institute of Technology North Bangkok, Department of Electrical Engineering; Mahidol University, Department of Research and Development Limwongse, Chanin; Mahidol University, Department of Research and Development Sinsomros, Saravudh; King Mongkut's Institute of Technology North Bangkok, Department of Electrical Engineering Yenchitsomanus, Pa-Thai; Mahidol University, Department of Research and Development Youngkong, Prakarnkiat; King Mongkut's Institute of Technology North Bangkok, Department of Electrical Engineering
Keywords:	Genetics, Pattern recognition

Variable-Length Haplotype Construction for Gene-Gene Interaction Studies

Anunchai Assawamakin[1] Nachol Chaiyaratana[1,2,*]

Chanin Limwongse[1] Saravudh Sinsomros[2]

Pa-Thai Yenchitsomanus[3] Prakarnkiat Youngkong[2]

November 1, 2007

Abstract

This paper presents a non-parametric classification technique for identifying a candidate bi-allelic genetic marker set that best describes disease susceptibility in gene-gene interaction studies. The developed technique functions by creating a mapping between inferred haplotypes and case/control status. The technique cycles through all possible marker combination models generated from the available marker set where the best interaction model is determined from prediction accuracy and two auxiliary criteria including low-to-high order haplotype propagation capability and model parsimony. Since variable-length haplotypes are created during the best model identification, the developed technique is referred to as a variable-length haplotype construction for gene-gene interaction (VarHAP) technique. VarHAP has been benchmarked against a multi-factor dimensionality reduction (MDR) program and a haplotype interaction technique

[1]Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Prannok Road, Bangkoknoi, Bangkok 10700, Thailand

[2]Department of Electrical Engineering, Faculty of Engineering, King Mongkut's Institute of Technology North Bangkok, 1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand

[3]Division of Medical Molecular Biology, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Prannok Road, Bangkoknoi, Bangkok 10700, Thailand

*Corresponding author, Tel: +66 2 913 2500 ext. 8410, Fax: +66 2 585 6149, E-mail: nchl@kmitnb.ac.th

embedded in a FAMHAP program in various two-locus interaction problems. The results reveal that VarHAP is suitable for all interaction situations with the presence of weak and strong linkage disequilibrium among genetic markers.

Keywords: Case-control studies; Gene-gene interaction; Haplotype; Linkage disequilibrium; Non-parametric classification

1 Introduction

Genetic epidemiology is a research field which aims to identify genetic polymorphisms that involve in disease susceptibility. Usual candidate polymorphisms include restriction fragment length polymorphisms (RFLPs), variable numbers of tandem repeats (VNTRs) and single nucleotide polymorphisms (SNPs). In recent years, SNPs are the most common choices due to simplicity and cost reduction in identification protocols. SNPs in diploid organisms are excellent bi-allelic genetic markers for various studies including genetic association, gene-gene interaction and gene-environment interaction. The availability of multiple SNPs on the same gene can also lead to haplotype analysis where genotypes of interest can be phased into pairs of haplotypes.

Traditional techniques for identification of relationship between a single SNP and disease susceptibility status involve various univariate statistical tests including χ^2 and odds ratio tests [1, 2]. However, many complementary computational techniques have been developed in the past decade to handle problems that involve multiple SNPs. Heidema et al. [3] have categorised these multi-locus techniques, which are capable of identifying a candidate SNP set from possible SNPs, into parametric and non-parametric methods. Examples of parametric method cover logistic regression techniques [4] and neural networks [5]. On the other hand, examples of non-parametric method include a set association approach [6], combinatorial techniques [7, 8, 9] and recursive partitioning techniques [10, 11]. In some of mentioned parametric [4, 5] and non-parametric [9] methods, pattern recognition and classification approaches have been successfully implemented as their core engines.

In addition to single and multiple SNP analysis, haplotype analysis has also gained attention from genetic epidemiologists. Haplotypes provide a record of evolutionary history more accurately than individual SNPs. Further, haplotypes can capture the patterns of linkage disequilibrium (LD)—a phenomenon where SNPs that are located in close proximity tend to travel together—in genome more accurately. Therefore, haplotypes may enable susceptibility gene identification in complex diseases more effectively than individual SNPs [12]. In lieu of this evidence, haplotype analysis should also be considered in addition to direct genotype analysis. Many computational techniques use haplotypes, which are inferred from multiple SNPs, as problem inputs. For instance, Sham et al. [13] proposes a logistic regression technique that produces a mapping model between haplotypes and disease status while Becker et al. [14] combine haplotype explanation probabilities of given genotypes from multiple gene or unlinked region data into a scalar statistic for a univariate test. Nonetheless, haplotypes have rarely been used as inputs for non-parametric classifiers for genetic association and interaction studies.

In this paper, a variable-length haplotype construction for gene-gene interaction (VarHAP) technique is proposed. The technique will involve non-parametric classification where haplotypes inferred from multiple SNP data are the classifier inputs. The chosen architecture for non-parametric classifier is the multifactor dimensionality reduction (MDR) technique [9]. Similar to the original MDR technique, the proposed technique would be able to identify appropriate candidate SNPs from possible SNPs and can be used in case-control genetic interaction studies. However, the technique would also be able to handle the situation where disease susceptibility is detectable in different haplotype backgrounds.

The organisation of this paper is as follows. In section 2, a brief explanation of MDR and the techniques for inferring haplotypes and obtaining haplotype explanation probability is given. The proposed VarHAP technique is then described in section 3. The test data and their description is given in section 4. Next, the results and discussions are described in section 5. Finally, the conclusions are drawn in section 6.

2 MDR, Haplotype Inference and Haplotype Explanation Probability

2.1 MDR

MDR is a classifier-based technique that is capable of identifying the best genetic marker combination among possible markers for the separation between case and control samples. Similar to other classification systems, a k-fold cross-validation technique provides a means to determine the classification accuracy of the candidate marker model. Basically, the combined case and control samples are randomly divided into k folds where $k-1$ folds of samples are used to construct a decision table for the classifier while the remaining fold of samples is used to identify the prediction capability of the constructed decision table. The decision table construction and testing procedure is repeated k times. Hence, the samples in each fold will always be utilised both to construct and to test the decision table. The number of cells in a decision table is given by G^{n_c} where n_c is the number of candidate markers selected from possible markers and G is the number of possible genotypes according to the marker. For a SNP, which is a bi-allelic marker, G is equal to three. During the decision table construction, each cell in the table is filled with case and control samples that have their genotype corresponds to the cell label. The ratio between numbers of case and control samples will provide the decision for each cell whether the corresponding genotype is a disease-predisposing or protective genotype. An example of decision table construction is illustrated in Figure 1. The prediction accuracy of the decision table is subsequently evaluated by counting the numbers of case and control samples in the testing fold that their disease status can be correctly identified using the constructed decision rules. The process of decision table construction and evaluation must be cycled through all or some of possible $2^{n_m}-1$ combinations where n_m is the total number of available markers in the study. The best genetic marker combination is determined from three criteria: prediction accuracy, cross-validation consistency and a sign test p-value. Each time that a testing fold is used

for prediction accuracy determination, the accuracy of the interested marker combination model can be compared with that from other models that also contain the same number of markers. The model that consistently ranks the first in comparison to other choices with the same amount of markers would have high cross-validation consistency. The non-parametric sign test p-value is calculated from the number of testing folds with accuracy greater than or equal to 50%. This single-tailed p-value is given by

$$p = \sum_{i=n_a}^{n_f} \binom{n_f}{i} \left(\frac{1}{2}\right)^{n_f} \tag{1}$$

where n_f is the total number of cross-validation folds and n_a is the number of cross-validation folds with testing accuracy $\geq 50\%$ [9]. Among three criteria, prediction accuracy is the main criterion for decision making while the other criteria are only used as auxiliary measures. Cross-validation consistency generally confirms that the high rank model can be consistently identified regardless of how the samples are divided for cross-validation. On the other hand, a sign test p-value indicates the number of testing folds with acceptable prediction accuracy and hence describes the usability of the model in the classification task. In the situation where two or more models with different number of markers are equally good in terms of prediction accuracy, cross-validation consistency and sign test p-value, the most parsimonious model—the combination with the least number of markers—will be the best model.

2.2 Haplotype Inference

With the availability of multiple SNPs from the same gene, haplotypes can be inferred from given genotypes. Let '0' and '1' denote the major (common) and minor (rare) alleles at a SNP location in a haplotype. A genotype can then be represented by a string, which consists of three characters: '0', '1' and '2'. In the genotype string, '0' denotes a homozygous wide-type site, '1' denotes a heterozygous site and '2' denotes a homozygous variant or homozygous mutant site. A genotype with all homozygous sites or single heterozygous site can

always be phased into one pair of haplotypes. On the other hand, a genotype with multiple heterozygous sites can be phased into multiple haplotype pairs. For example, genotype `0102` leads to haplotypes `0001` and `0101` while genotype `0112` leads to two possible haplotype pairs: `0001/0111` and `0011/0101`. Many algorithms exist for haplotype inference [15, 16, 17]. In this paper, an expectation-maximisation algorithm [15] is the chosen technique due to its simplicity and implementation efficacy. Regardless of the inference technique employed, the usual result from an inference algorithm covers haplotype frequencies and possible haplotype phases of each genotype.

2.3 Haplotype Explanation Probability

In a genomic region with multiple heterozygous sites, multiple pairs of haplotypes can be inferred from a given genotype. The probability of a genotype to be phased into one specific pair of haplotypes would depend on the frequencies of haplotypes constituting the pairs [14]. This probability is given by

$$w_{ij} = \frac{f_i f_j}{\sum_{(h_k,h_l)\in H} f_k f_l} \tag{2}$$

where w_{ij} is the probability for haplotype pair ij, f_i denotes the frequency of the ith haplotype, h_k is the kth haplotype and H represents the set of haplotype explanations which are compatible with the genotype of interest. For example, genotype `0110` can be phased into two haplotype pairs: `0000/0110` (h_1/h_4) and `0010/0100` (h_2/h_3). If the frequencies for haplotypes `0000`, `0010`, `0100` and `0110` are respectively 0.5, 0.2, 0.2 and 0.1, the probabilities for the pairs `0000/0110` and `0010/0100` are 0.556 and 0.444. Obviously, the probability of a genotype with all homozygous sites or single heterozygous site to be phased into a pair of haplotypes would be equal to one. In genetic interaction studies where the number of genes or unlinked regions is greater than one, the haplotype explanation probabilities from all regions can be combined together. An overall contribution by one sample to haplotype

configuration $(h_j^1, h_j^2, \ldots, h_j^{n_u})$ in a study with n_u genes/unlinked regions is given by

$$c_{(h_j^1, h_j^2, \ldots, h_j^{n_u})} = 2 \prod_{i=1}^{n_u} w_{jk}^i \frac{(1 + \delta_{jk}^i)}{2} \tag{3}$$

where $c_{(h_j^1, h_j^2, \ldots, h_j^{n_u})}$ is the contribution value and δ is defined as $\delta_{jk} = 1$ for $j = k$ and $\delta_{jk} = 0$ for $j \neq k$. In the previous example where haplotypes from only one region are considered, $c_{h_1} = 0.556$, $c_{h_2} = 0.444$, $c_{h_3} = 0.444$ and $c_{h_4} = 0.556$. Notice that the sum of contribution values is equal to two; this reflects the fact that each genotype is made up from two haplotypes. Becker et al. [14] use this contribution value in the construction of a contingency table where a χ^2 test statistic is subsequently calculated. With the use of a Monte Carlo simulation, an estimated p-value is then obtained for the test statistic. Similar to the model exploration strategy in MDR, the process of contingency table construction and p-value calculation can also be cycled through all or some of possible interaction models. The model with appropriate candidate SNPs taken from possible SNPs is the one with minimum p-value and is said to be the best model for interaction explanation. This statistics-based procedure can be found as an integral part of the FAMHAP program [18].

3 VarHAP

VarHAP is proposed for case-control interaction studies. Similar to MDR, the technique is also a classifier-based technique. However, instead of using a genotype data analysis as a means to identify the best SNP combination, the decision table for classification is constructed from the haplotype contribution value described earlier. As a result, haplotypes with different lengths must be inferred during the search for the best model. The number of decision cells during the consideration on haplotypes constructed from a specific set of SNPs is governed by the total number of possible haplotype configurations as illustrated in Figure 2. In brief, VarHAP would maintain the ability to find the best SNP combination while also be able to identify possible disease-predisposing and protective haplotype configurations.

Since VarHAP is essentially a classification system, the principal criterion for choosing the optimal SNP combination model is still the prediction accuracy. However, with the use of haplotype contribution value as a means for decision rule construction, an additional model selection criterion that exploits the nature of haplotype can be formulated. This criterion can be referred to as haplotype propagation capability. Basically, if a haplotype constructed from a specific set of SNPs is related to disease susceptibility status, haplotypes constructed from a SNP set which is a superset of the previously specified SNPs should also predict the same relationship. This implies that predisposing and protective haplotypes in a low-order model must be able to propagate into haplotypes in high-order models. For example, consider a single-gene problem with four possible SNPs: `X1`, `X2`, `X3` and `X4`. If haplotypes in the model with SNPs `(X2, X4)` are related to disease susceptibility, haplotypes in the models with SNPs `(X1, X2, X4)`, `(X2, X3, X4)` and `(X1, X2, X3, X4)` should produce the same result. The haplotype propagation capability, which is a dichotomous criterion, can be determined from the evidence that the sign test p-value and the prediction accuracy can be maintained throughout the process of model order increment. Again, in the situation where two or more models with different number of SNPs are equally good in terms of both prediction accuracy and haplotype propagation capability, the most parsimonious model will be the best model.

4 Data Sets

The performance of the proposed VarHAP technique is evaluated through benchmark trials. 12 simulated data sets, which represent various gene-gene interaction phenomena including epistasis and heterogeneity, are considered [14, 19]. Each data set contains 600 case samples and 600 control samples. Each sample consists of 10 total SNPs from two genes where five SNPs exist in each gene. All SNPs in control samples are in Hardy-Weinberg equilibrium [20]. Only one SNP from each gene is interacted with one another. The two-

locus interaction models are illustrated in Table 1. The epistatic models Ep-1–Ep-6 and the heterogeneity models Het-1–Het-3 have been discussed by Neuman and Rice [21], who also provide examples of diseases for which these models may be applicable. The heterogeneity models S-1 and S-2 and the epistatic model S-3 have been investigated by Schork et al. [22]. From Table 1, if the frequency of the disease allele at a locus is greater than 0.5, the major allele is the disease allele. Otherwise, the minor allele is the disease allele. These interaction models describe disease susceptibility status in terms of penetrance. Penetrance of a genotype with a specific number of disease alleles is the probability that a subject with this genotype has the disease. The test data sets are simulated by a genomeSIM package [23] with the default setting. As a result, it is also possible to vary the LD pattern among SNPs in the same gene. This leads to two main case studies that need to be explored: strong LD and weak LD cases. In the strong LD case, the susceptibility-causative SNP in each gene and its two adjacent SNPs are in linkage disequilibrium where Lewontin's D' value [24] is in the range of 0.80–0.95. In contrast, the Lewontin's D' value for each pairwise LD measurement between susceptibility-causative SNP and its adjacent SNPs is in the range of 0.50–0.60 in the weak LD case. In the strong LD case, an interaction detection technique should be able to identify both the actual two-locus model that directly leads to disease susceptibility and other alternative models which consist of SNPs in strong LD patterns. The ability to detect these other models is important. This is because it is not always straightforward to identify SNPs which are responsible for disease susceptibility in real case-control interaction studies. In contrast, an interaction detection technique should narrow the search to the original two-locus model in weak LD case since it is the only usable model.

5 Results and Discussions

VarHAP is benchmarked against MDR and FAMHAP. Since the test data contains 10 SNPs, all three techniques have to explore $2^{10} - 1 = 1,023$ possible SNP combination models. An

initial investigation reveals that with the use of minimum p-value as the sole model selection criterion, FAMHAP reports a large number of models with the estimated p-value equals to zero. As a result, haplotype propagation capability is also implemented as an additional model selection criterion. Further, the parsimony criterion is also utilised when there is a tie between multiple models with different number of SNPs. The results from all three techniques in weak and strong LD case studies are summarised in Tables 2 and 3, respectively.

The prediction accuracy of MDR is higher than that of VarHAP in both case studies. This is because VarHAP uses contribution values which are obtained from inferred haplotypes instead of inferred diplotypes—pairs of haplotypes that together describe correct phases of given genotypes—to create decision rules. Consider a situation where disease susceptibility can be determined from a single SNP where the predisposing genotype is the homozygous variant. In other words, the disease susceptibility can be described by a recessive genetic model. MDR can easily classify the heterozygous and homozygous wide-type genotypes as protective genotypes. However, VarHAP would only correctly classify both homozygous genotypes since each genotype is made up from two copies of the same haplotype: two major alleles for the homozygous wide-type and two minor alleles for the homozygous variant. VarHAP would partially misclassify samples with heterozygous genotype. This is because VarHAP identifies the major allele as the protective allele and the minor allele as the predisposing allele. In order to increase the prediction accuracy of VarHAP, it may be necessary to construct decision tables from diplotype information instead of haplotype contribution values. Nonetheless, this will also rapidly increase the dimensions of decision tables in VarHAP.

In the weak LD case study, both MDR and VarHAP are able to identify correct sets of SNPs that lead to disease susceptibility. On the other hand, FAMHAP reports both actual and alternative interaction models. This is undesirable since it would not be possible to further explain disease susceptibility from multiple candidate models in the absence of strong linkage disequilibrium among SNPs. In other words, FAMHAP is quite sensitive in

this situation. Further analysis reveals that MDR is marginally better than VarHAP in two epistasis problems: Ep-2 and Ep-5. MDR correctly identifies models which contain two SNPs while the models located by VarHAP contain a few extra SNPs. Nonetheless, these two models identified by VarHAP are still useful to susceptibility explanation.

All three techniques are able to locate correct interaction models in the strong LD case study. However, only FAMHAP and VarHAP are capable of identifying alternative models. Since MDR suggests one candidate model for each fixed-number SNP set, it would not be possible for MDR to produce any alternative models. Recall that these alternative models are equally important since SNPs in the principal two-locus interaction model and SNPs from an alternative model are in strong linkage disequilibrium. This implies that disease susceptibility can be explained using either the original interaction model or the alternative models. This disadvantage in MDR can be overcome if the cross-validation consistency criterion can be replaced by other decision criteria. In this case study, FAMHAP is marginally better than VarHAP in terms of alternative model identification in three epistasis and heterogeneity problems: Ep-3, Ep-6 and Het-3. This means that FAMHAP is at its best when SNPs are in strong linkage disequilibrium. Nonetheless, the overall results from both case studies suggest that VarHAP is the best technique. This is concluded from the fact that VarHAP does not report ambiguous results in weak LD case study while is also capable of producing alternative models in strong LD case study. This is crucial because it is impossible to know beforehand whether susceptibility-causative SNPs are in weak or strong linkage disequilibrium with other SNPs in real case-control interaction studies. In other words, a technique that performs satisfactorily in both weak and strong LD cases would have an advantage over a technique that functions well in only one scenario.

6 Conclusions

In this paper, a non-parametric pattern recognition/classification technique for case-control gene-gene interaction studies is presented. Instead of using direct genotype inputs in classification, inferred haplotypes, which are obtained through an expectation-maximisation algorithm [15], are used as inputs. Each case/control sample contributes values derived from inferred haplotypes to decision tables which are constructed and tested for all possible gene-gene interaction models. The technique primarily uses prediction accuracy obtained from k-fold cross-validation as a means for identifying candidate SNPs which are responsible for disease susceptibility. The technique also employs haplotype propagation capability as an additional criterion. If the selection procedure ends in a tie between two or more models with different number of SNPs, the most parsimonious model is then reported as the interaction model. Since haplotypes with different length must be constructed during model identification, the proposed technique can be referred to as a variable-length haplotype construction for gene-gene interaction (VarHAP) technique. VarHAP has been benchmarked against two interaction model detection programs namely MDR [9] and FAMHAP [14, 18] in 12 two-locus epistasis and heterogeneity problems [14, 19]. The results reveal that FAMHAP reports multiple ambiguous models in the presence of weak linkage disequilibrium among input SNPs while MDR is not suitable for alternative interaction model identification when input SNPs are in strong linkage disequilibrium. In contrast, VarHAP emerges as the most suitable technique in both situations involving weak and strong linkage disequilibrium. Suggestions for further improvement of MDR and VarHAP are also included.

Supplementary Information

VarHAP, which is implemented in Java, and the simulated data sets used in the article are available upon request (`e-mail: nchl@kmitnb.ac.th`). In addition to the use of the genomeSIM package [23], the data sets can also be generated by a SNaP package [25]. Readers

might also be interested in applying the techniques discussed in this article to examples of case-control data sets, which are publicly available from the Wellcome Trust Case Control Consortium [26].

Acknowledgments

Anunchai Assawamakin was supported by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Programme (Grant No. PHD/4.I.MU.45/C.1) and the National Center for Genetic Engineering and Biotechnology (BIOTEC), the National Science and Technology Development Agency (NSTDA). Nachol Chaiyaratana was supported by the Thailand Research Fund and the Commission on Higher Education (CHE). Prakarnkiat Youngkong was supported by the Commission on Higher Education. The authors acknowledge Scott M. Dudek at the Vanderbilt University for providing an access to the genomeSIM package.

References

[1] C.M. Lewis, "Genetic association studies: Design, analysis and interpretation," *Briefings in Bioinformatics*, vol. 3, pp. 146–153, June 2002.

[2] G. Montana, "Statistical methods in genetics," *Briefings in Bioinformatics*, vol. 7, pp. 297–308, Sep. 2006.

[3] A.G. Heidema, J.M.A. Boer, N. Nagelkerke, E.C.M. Mariman, D.L. Van der A, and E.J.M. Feskens, "The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases," *BMC Genetics*, vol. 7, pp. 23, Apr. 2006.

[4] N. Nagelkerke, J. Smits, S. Le Cessie, and H. Van Houwelingen, "Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting," *Statistics in Medicine*, vol. 24, 121–130, Jan. 2005.

[5] M.D. Ritchie, B.C. White, J.S. Parker, L.W. Hahn, and J.H. Moore, "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases," *BMC Bioinformatics*, vol. 4, pp. 28, July 2003.

[6] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies," *Genome Research*, vol. 11, pp. 2115–2119, Dec. 2001.

[7] M.R. Nelson, S.L.R. Kardia, R.E. Ferrell, and C.F. Sing, "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation," *Genome Research*, vol. 11, pp. 458–470, Mar. 2001.

[8] R. Culverhouse, T. Klein, and W. Shannon, "Detecting epistatic interactions contributing to quantitative traits," *Genetic Epidemiology*, vol. 27, pp. 141–152, Sep. 2004.

[9] L.W. Hahn, M.D. Ritchie, and J.H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, pp. 376–382, Feb. 2003.

[10] K.L. Lunetta, L.B. Hayward, J. Segal, and P. Van Eerdewegh, "Screening large-scale association study data: Exploiting interactions using random forests," *BMC Genetics*, vol. 5, pp. 32, Dec. 2004.

[11] A. Bureau, J. Dupuis, K. Falls, K.L. Lunetta, B. Hayward, T.P. Keith, and P. Van Eerdewegh, "Identifying SNPs predictive of phenotype using random forests," *Genetic Epidemiology*, vol. 28, pp. 171–182, Feb. 2005.

[12] E.K. Silverman, "Haplotype thinking in lung disease," *Proceedings of the American Thoracic Society*, vol. 4, pp. 4–8, Jan. 2007.

[13] P.C. Sham, F.V. Rijsdijk, J. Knight, A. Makoff, B. North, and D. Curtis, "Haplotype association analysis of discrete and continuous traits using mixture of regression models," *Behavior Genetics*, vol. 34, pp. 207–214, Mar. 2004.

[14] T. Becker, J. Schumacher, S. Cichon, M.P. Baur, and M. Knapp, "Haplotype interaction analysis of unlinked regions," *Genetic Epidemiology*, vol. 29, pp. 313–322, Dec. 2005.

[15] L. Excoffier and M. Slatkin, "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population," *Molecular Biology and Evolution*, vol. 12, pp. 921–927, Sep. 1995.

[16] M. Stephens, N.J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics*, vol. 68, pp. 978–989, Apr. 2001.

[17] T. Niu, Z.S. Qin, X. Xu, and J.S. Liu, "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms," *American Journal of Human Genetics*, vol. 70, pp. 157–169, Jan. 2002.

[18] T. Becker and M. Knapp, "A powerful strategy to account for multiple testing in the context of haplotype analysis," *American Journal of Human Genetics*, vol. 75, pp. 561–570, Oct. 2004.

[19] M. Knapp, S.A. Seuchter, and M.P. Baur, "Two-locus disease models with two marker loci: The power of affected-sib-pair tests," *American Journal of Human Genetics*, vol. 55, pp. 1030–1041, Nov. 1994.

[20] G.H. Hardy, "Mendelian proportions in a mixed population," *Science*, vol. 28, pp. 49–50, July 1908.

[21] R.J. Neuman and J.P. Rice, "Two-locus models of disease," *Genetic Epidemiology*, vol. 9, pp. 347–365, 1992.

[22] N.J. Schork, M. Boehnke, J.D. Terwilliger, and J. Ott, "Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits," *American Journal of Human Genetics*, vol. 53, pp. 1127–1136, Nov. 1993.

[23] S.M. Dudek, A.A. Motsinger, D.R. Velez, S.M. Williams, and M.D. Ritchie, "Data simulation software for whole-genome association and other studies in human genetics," in *Pacific Symposium on Biocomputing 2006*, R.B. Altman, A.K. Dunker, L. Hunter, T. Murray, and T.E. Klein, eds., Singapore: World Scientific, 2006, pp. 499–510.

[24] R.C. Lewontin, "On measures of gametic disequilibrium," *Genetics*, vol. 120, pp. 849–852, Nov. 1988.

[25] M. Nothnagel, "Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods," *American Journal of Human Genetics*, vol. 71 (Suppl.), pp. A2363, Oct. 2002.

[26] Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, June 2007.

Biography

Anunchai Assawamakin is a Ph.D. student at Mahidol University. He also received his B.Sc. degree from Mahidol University. His current research interests include human genetics and bioinformatics.

Nachol Chaiyaratana is an associate professor of electrical engineering at King Mongkut's Institute of Technology North Bangkok and an adjunct professor of medicine at Mahidol University. He received his B.Eng. and Ph.D. degrees from the University of Sheffield. His current research interests include evolutionary computation, machine learning and bioinformatics.

Chanin Limwongse is the Head of Division of Molecular Genetics at the Department of Research and Development, Mahidol University. He also received his M.D. degree from Mahidol University. His current research interests include complex diseases and human genetics.

Saravudh Sinsomros is an M.Eng. student at King Mongkut's Institute of Technology North Bangkok. He received his B.Eng. degree from Thammasat University. His current research interests include machine learning and bioinformatics.

Pa-Thai Yenchitsomanus is a professor of medicine and the Head of Division of Medical Molecular Biology at the Department of Research and Development, Mahidol University. He received his B.Sc., M.S. and Ph.D. degrees from Chiang Mai University, Mahidol University and Australian National University, respectively. His current research interests include complex diseases and molecular genetics.

Prakarnkiat Youngkong is a Ph.D. student at King Mongkut's Institute of Technology North Bangkok. He received his B.S. and M.S. degrees from Rensselaer Polytechnic Institute and University of California Los Angeles, respectively. His current research interests include machine learning and bioinformatics.

List of Figures

List of Tables

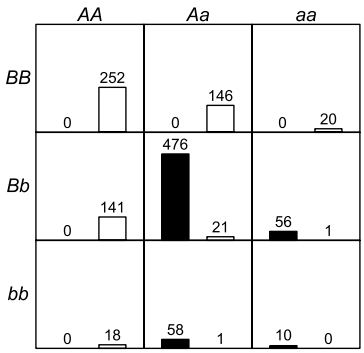

Figure 1: An MDR decision table which is constructed using 1,200 case-control samples. The genotype of each sample is determined from two SNPs. The table consists of nine cells where each cell represents a unique genotype. The left (black) bar in each cell represents the number of case samples while the right (white) bar represents the number of control samples. The cells with genotypes *AaBb*, *aaBb*, *Aabb* and *aabb* are labelled as predisposing genotypes while the cells with genotypes *AABB*, *AaBB*, *aaBB*, *AABb* and *AAbb* are labelled as protective genotypes.

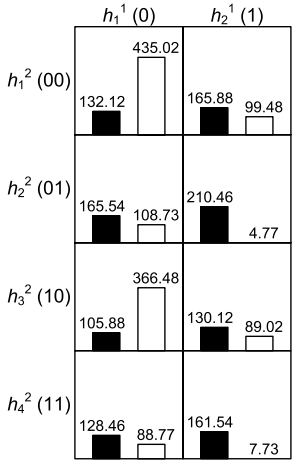

Figure 2: A VarHAP decision table which is constructed from 1,200 case-control samples. Haplotypes in the first gene are obtained from one SNP while haplotypes in the second gene are inferred from two SNPs. The table consists of eight cells where each cell represents a unique haplotype configuration. The left (black) bar in each cell represents the accumulative contribution from case samples while the right (white) bar represents the accumulative contribution from control samples. The cells with haplotype configurations (h_2^1, h_1^2), (h_1^1, h_2^2), (h_2^1, h_2^2), (h_2^1, h_3^2), (h_1^1, h_4^2) and (h_2^1, h_4^2) are labelled as predisposing haplotype configurations while the cells with haplotype configurations (h_1^1, h_1^2) and (h_1^1, h_3^2) are labelled as protective haplotype configurations.

Table 1: Description of two-locus disease models. d_{ij} is the penetrance of a genotype carrying i disease alleles at locus 1 and j disease alleles at locus 2. p_1 is the frequency of the disease allele at locus 1 while p_2 is the frequency of the disease allele at locus 2. $\psi = 2\phi - \phi^2$.

Model	d_{22}	d_{21}	d_{20}	d_{12}	d_{11}	d_{10}	d_{02}	d_{01}	d_{00}	p_1	p_2	ϕ
Ep-1	ϕ	ϕ	0	ϕ	ϕ	0	0	0	0	0.210	0.210	0.707
Ep-2	ϕ	ϕ	0	0	0	0	0	0	0	0.600	0.199	0.778
Ep-3	ϕ	0	0	0	0	0	0	0	0	0.577	0.577	0.900
Ep-4	ϕ	ϕ	0	ϕ	0	0	ϕ	0	0	0.372	0.243	0.911
Ep-5	ϕ	ϕ	0	ϕ	0	0	0	0	0	0.349	0.349	0.799
Ep-6	0	ϕ	ϕ	ϕ	0	0	ϕ	0	0	0.190	0.190	1.000
Het-1	ψ	ψ	ϕ	ψ	ψ	ϕ	ϕ	ϕ	0	0.053	0.053	0.495
Het-2	ψ	ψ	ϕ	ϕ	ϕ	0	ϕ	ϕ	0	0.279	0.040	0.660
Het-3	ψ	ϕ	ϕ	ϕ	0	0	ϕ	0	0	0.194	0.194	1.000
S-1	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	0	0.052	0.052	0.522
S-2	1	1	1	ϕ	ϕ	0	ϕ	ϕ	0	0.228	0.045	0.574
S-3	1	1	ϕ	1	ϕ	0	ϕ	0	0	0.194	0.194	0.512

Table 2: MDR, VarHAP and FAMHAP results from the weak LD case study. 10-fold cross-validation is used in MDR and VarHAP. The prediction accuracy is obtained for the identified principal interaction model. Estimated p-values in FAMHAP results are equal to zero while sign test p-values in MDR and VarHAP results are less than 0.001 in all two-locus problems. The technique is said to be able to identify the correct gene-gene interaction model if the reported principal model contains both SNPs which are directly participated in the interaction model. Alternative models are models which contain at least two SNPs where each SNP must be either a SNP from the two-locus model or a SNP which is in linkage disequilibrium with one of the SNPs from the model. The number in each bracket denotes the order of the identified model (the number of SNPs in the model).

Two-Locus Model	MDR Prediction Accuracy (%)	VarHAP Prediction Accuracy (%)	Correct Model Identification Technique	Alternative Model Identification Technique
Ep-1	98.00	73.92	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-2	98.58	78.39	MDR(2), VarHAP(4), FAMHAP(2)	FAMHAP(2)
Ep-3	99.50	87.50	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-4	99.25	78.96	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-5	98.42	75.19	MDR(2), VarHAP(3), FAMHAP(2)	FAMHAP(2)
Ep-6	100.00	85.10	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Het-1	93.75	73.29	MDR(2), VarHAP(2), FAMHAP(2)	
Het-2	97.33	78.40	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Het-3	100.00	84.40	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
S-1	94.00	72.98	MDR(2), VarHAP(2), FAMHAP(2)	
S-2	97.58	79.81	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
S-3	96.75	79.15	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)

Table 3: MDR, VarHAP and FAMHAP results from the strong LD case study. The explanation for how the results are obtained and displayed is the same as that given in Table 2.

Two-Locus Model	MDR Prediction Accuracy (%)	VarHAP Prediction Accuracy (%)	Correct Model Identification Technique	Alternative Model Identification Technique
Ep-1	98.00	73.92	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
Ep-2	98.58	77.02	MDR(2), VarHAP(4), FAMHAP(2)	VarHAP(4), FAMHAP(2)
Ep-3	99.50	87.50	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Ep-4	99.25	78.96	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
Ep-5	98.42	75.87	MDR(2), VarHAP(3), FAMHAP(2)	VarHAP(3), FAMHAP(2)
Ep-6	100.00	85.10	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
Het-1	93.75	75.41	MDR(2), VarHAP(3), FAMHAP(2)	VarHAP(3), FAMHAP(2)
Het-2	97.33	78.40	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
Het-3	100.00	84.40	MDR(2), VarHAP(2), FAMHAP(2)	FAMHAP(2)
S-1	94.00	72.98	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
S-2	97.58	79.81	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)
S-3	96.75	79.15	MDR(2), VarHAP(2), FAMHAP(2)	VarHAP(2), FAMHAP(2)

Reply to Comments

Dear Guest Editor,

Thank you very much for your invitation to include this article in a special issue of IEEE Engineering in Medicine and Biology Magazine. The article has been revised according to all reviewers' comments. Since the length of the revised article is limited to five magazine pages, the revised article needs to be concise. A summary of the revision is listed (in italic) below.

Both reviewers noted the lack of experiments on real data and comparisons. I recommend authors adequately address the reviewers' comments.

The proposed VarHAP technique has been successfully benchmarked against MDR (Hahn et al., 2003) and FAMHAP (Becker and Knapp, 2004; Becker et al., 2005). Detailed explanations about MDR, FAMHAP and VarHAP have been given in sections 2.1, 2.3 and 3, respectively. An explanation regarding the possibility of using the proposed technique on real case-control data sets has also been given.

References

T. Becker and M. Knapp, "A powerful strategy to account for multiple testing in the context of haplotype analysis," American Journal of Human Genetics, vol. 75, pp. 561–570, Oct. 2004.

T. Becker, J. Schumacher, S. Cichon, M.P. Baur, and M. Knapp, "Haplotype interaction analysis of unlinked regions," Genetic Epidemiology, vol. 29, pp. 313–322, Dec. 2005.

L.W. Hahn, M.D. Ritchie, and J.H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," Bioinformatics, vol. 19, pp. 376–382, Feb. 2003.

Reply to Reviewer 1

Dear Reviewer 1,

We are very grateful to receive your valuable and helpful comments. They have been considered and taken as the main guideline to revise our paper in order to upgrade its quality. Here, we kindly attached our explanations corresponding to your comments and queries. They are listed (in italic) below.

Interesting new method, good English, literature covered well.

To improve:

- need more details on comparison benchmarks with other methods for their evaluation criteria for same datasets

The proposed VarHAP technique has been benchmarked against MDR (Hahn et al., 2003) and FAMHAP (Becker and Knapp, 2004; Becker et al., 2005). All three techniques share the same strategy where multiple susceptibility explanation models are examined. These models are created from various combinations of genetic markers extracted from available markers in case-control data. However, MDR takes genotype information as inputs while VarHAP and FAMHAP uses inferred haplotypes as inputs. Both MDR and VarHAP are non-parametric classifier-based techniques. As a result, their principal evaluation criterion for selecting the best susceptibility explanation model is the prediction accuracy. On the other hand, FAMHAP is a statistics-based technique. Hence, its principal evaluation criterion is the global p-value, which is estimated via a Monte Carlo simulation. The VarHAP performance is compared with that of MDR and FAMHAP since VarHAP essentially uses MDR classification engine while takes inferred haplotype inputs in a similar manner to FAMHAP. Consequently, the technique is able to handle the situation where disease susceptibility is detectable in different haplotype backgrounds. Detailed explanations about MDR, FAMHAP and VarHAP have been given in sections 2.1, 2.3 and 3, respectively.

- add details about simulated datasets, so others can replicate the datasets used

Details about the simulated data sets have been extended and included in section 4. These

simulated data sets are available from the corresponding author upon request as indicated in the supplementary information section. With the use of two-locus disease models given in Table 1, the case-control data sets can also be generated by a SNaP package (Nothnagel, 2002), which is available from `http://capella.uni-kiel.de/snap/snap.htm`.

- please indicate availability/implementation of the software

VarHAP has been implemented in Java. The program is also available upon request.

References

T. Becker and M. Knapp, "A powerful strategy to account for multiple testing in the context of haplotype analysis," American Journal of Human Genetics, vol. 75, pp. 561–570, Oct. 2004.

T. Becker, J. Schumacher, S. Cichon, M.P. Baur, and M. Knapp, "Haplotype interaction analysis of unlinked regions," Genetic Epidemiology, vol. 29, pp. 313–322, Dec. 2005.

L.W. Hahn, M.D. Ritchie, and J.H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," Bioinformatics, vol. 19, pp. 376–382, Feb. 2003.

M. Nothnagel, "Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods," American Journal of Human Genetics, vol. 71 (Suppl.), pp. A2363, Oct. 2002.

Reply to Reviewer 2

Dear Reviewer 2,

We are very grateful to receive your valuable and helpful comments. They have been considered and taken as the main guideline to revise our paper in order to upgrade its quality. Here, we kindly attached our explanations corresponding to your comments and queries. They are listed (in italic) below.

The method is tested on a simulated datasets. Any reason for not using real datasets?

Methodologies in genetic epidemiology are usually benchmarked via the use of simulated

data. This is because the data can be easily generated for various disease susceptibility scenarios. Real case-control data sets are often proven to be unsuitable for multi-scenario benchmarking due to the prohibitively large expenses for data collection and the lack of complete knowledge regarding susceptibility models for the diseases of interest. Nonetheless, inclusion of results interpreted from a few real data sets in methodology literature is quite common. For instance, a real case-control data set, which contains multiple SNPs on the same gene or unlinked region and is suitable for the proposed VarHAP technique, has been described in Becker et al. (2005). The data set consists of ten SNPs where the first seven SNPs are located on the D-amino acid oxidase activator (DAOA, formerly known as G72) gene while the three remaining SNPs are from the D-amino acid oxidase (DAAO) gene. Associations between these two genes and schizophrenia have been reported by Chumakov et al. (2002) and Schumacher et al. (2004). The first author of the manuscript has made a request for this data set to the corresponding author of Becker et al. (2005). Unfortunately, the request has been denied due to the necessity of a further investigation into the disease susceptibility by the authors of Becker et al. (2005). The authors of this manuscript are also aware of publicly available case-control data sets from the Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium, 2007). Since the data sets contain genome-wide SNP information, the process of selecting suitable SNPs from multiple candidate genes would take longer than the time allowed for the manuscript revision (30 days). It is noted that these genome-wide data sets are publicly available after the first review process has been completed.

The authors have addressed the reviewers comments sufficiently and the paper is ready for acceptence.

Thank you for your kind comments. The paper has also been revised according to the comments from the second review.

References

T. Becker, J. Schumacher, S. Cichon, M.P. Baur, and M. Knapp, "Haplotype interaction analysis of unlinked regions," Genetic Epidemiology, vol. 29, pp. 313–322, Dec. 2005.

I. Chumakov, M. Blumenfeld, O. Guerassimenko, L. Cavarec, M. Palicio, H. Abderrahim, L. Bougueleret, C. Barry, H. Tanaka, P. La Rosa, A. Puech, N. Tahri, A. Cohen-Akenine, S. Delabrosse, S. Lissarrague, F.P. Picard, K. Maurice, L. Essioux, P. Millasseau, P. Grel, V. Debailleul, A.M. Simon, D. Caterina, I. Dufaure, K. Malekzadeh, M. Belova, J.J. Luan, M. Bouillot, J.L. Sambucy, G. Primas, M. Saumier, N. Boubkiri, S. Martin-Saumier, M. Nasroune, H. Peixoto, A. Delaye, V. Pinchot, M. Bastucci, S. Guillou, M. Chevillon, R. Sainz-Fuertes, S. Meguenni, J. Aurich-Costa, D. Cherif, A. Gimalac, C. Van Duijn, D. Gauvreau, G. Ouellette, I. Fortier, J. Raelson, T. Sherbatich, N. Riazanskaia, E. Rogaev, P. Raeymaekers, J. Aerssens, F. Konings, W. Luyten, F. Macciardi, P.C. Sham, R.E. Straub, D.R. Weinberger, N. Cohen, and D. Cohen, "Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia," Proceedings of the National Academy of Sciences of the United States of America, vol. 99, pp. 13675–13680, Oct. 2002.

J. Schumacher, R. Abon Jamra, J. Freudenberg, T. Becker, S. Ohlraun, A.C.J. Otte, M. Tullius, S. Kovalenko, A. Van Den Bogaert, W. Maier, M. Rietschel, P. Propping, M.M. Nöthen, and S. Cichon, "Examination of G72 and D-amino-acid oxidase as genetic risk factors for schizophrenia and bipolar affective disorder," Molecular Psychiatry, vol. 9, pp. 203–207, Feb. 2004.

Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," Nature, vol. 447, pp. 661–678, June 2007.

Department of Electrical Engineering
Faculty of Engineering
King Mongkut's Institute of Technology North Bangkok
1518 Piboolsongkram Road, Bangsue, Bangkok 10800, Thailand
Tel: +66 2 9132500 Ext. 8410
Fax: +66 2 5856149
E-mail: nchl@kmitnb.ac.th

1 November 2007

Dear Prof. Jagath C. Rajapakse,

The attached files contain a revised manuscript entitled "Variable-Length Haplotype Construction for Gene-Gene Interaction Studies" by Assawamakin et al. and a reply to comments. The manuscript is submitted for publication in *IEEE Engineering in Medicine and Biology Magazine*. The paper describes an attempt to develop an automated decision support tool, which can be used in the detection of gene-gene interaction for disease susceptibility explanation.

On behalf of all authors, I guarantee that this paper has not been published previously and is not under consideration for publication elsewhere.

Yours Sincerely,

Dr. Nachol Chaiyaratana

References

Becker, T. and Knapp, M. (2004). A powerful strategy to account for multiple testing in the context of haplotype analysis. *American Journal of Human Genetics*, 75, 561–570.

Becker, T., Schumacher, J., Cichon, S., Baur, M. P. and Knapp, M. (2005). Haplotype interaction analysis of unlinked regions. *Genetic Epidemiology*, *29*, 313–322.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, *28*, 171–182.

Culverhouse, R., Klein, T. and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology*, *27*, 141–152.

Dudek, S. M., Motsinger, A. A., Velez, D. R., Williams, S. M. and Ritchie, M. D. (2006). Data simulation software for whole-genome association and other studies in human genetics. In R. B. Altman, A. K. Dunker, L. Hunter, T. Murray and T. E. Klein (eds.), *Pacific Symposium on Biocomputing 2006*. Singapore: World Scientific. pp. 499–510.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, *12*, 921–927.

Hahn, L. W., Ritchie, M. D. and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, *19*, 376–382.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, *28*, 49–50.

Heidema, A. G., Boer, J. M. A., Nagelkerke, N., Mariman, E. C. M., Van der A, D. L. and Feskens, E. J. M. (2006). The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics*, *7*, 23.

Hoh, J., Wille, A. and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research*, *11*, 2115–2119.

Knapp, M., Seuchter, S. A. and Baur, M. P. (1994). Two-locus disease models with two marker loci: The power of affected-sib-pair tests. *American Journal of Human Genetics*, *55*, 1030–1041.

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, *3*, 146–153.

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics*, *120*, 849–852.

Lunetta, K. L., Hayward, L. B., Segal, J. and Van Eerdewegh, P. (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics*, *5*, 32.

Montana, G. (2006). Statistical methods in genetics. *Briefings in Bioinformatics*, *7*, 297–308.

Nagelkerke, N., Smits, J., Le Cessie, S. and Van Houwelingen, H. (2005). Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting. *Statistics in Medicine*, *24*, 121–130.

Nelson, M. R., Kardia, S. L. R., Ferrell, R. E. and Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, *11*, 458–470.

Neuman, R. J. and Rice, J. P. (1992). Two-locus models of disease. *Genetic Epidemiology*, *9*, 347–365.

Niu, T., Qin, Z. S., Xu, X. and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, *70*, 157–169.

Nothnagel, M. (2002). Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. *American Journal of Human Genetics*, *71*(Suppl.), A2363.

Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W. and Moore, J. H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, *4*, 28.

Schork, N. J., Boehnke, M., Terwilliger, J. D. and Ott, J. (1993). Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits. *American Journal of Human Genetics*, *53*, 1127–1136.

Sham, P. C., Rijsdijk, F. V., Knight, J., Makoff, A., North, B. and Curtis, D. (2004). Haplotype association analysis of discrete and continuous traits using mixture of regression models. *Behavior Genetics*, *34*, 207–214.

Silverman, E. K. (2007). Haplotype thinking in lung disease. *Proceedings of the American Thoracic Society*, *4*, 4–8.

Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, *68*, 978–989.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661–678.