Machine Learning to Interrogate High-throughput Genomic Data: Theory and Applications

Guoqiang Yu

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Electrical Engineering

Yue Wang, Chair

Jianhua Xuan, Co-chair

William T. Baumann, Member

Chang-Tien Lu, Member

Ge Wang, Member

Robert Clarke, Member

September 7, 2011

Arlington, Virginia

Keywords: Genome-wide Association Study, Multi-category gene selection, Gene-Gene Interaction, Gene-Environment Interaction.

Biographical Sketch

Guoqiang Yu is a doctoral graduate student in the Bradley Department of Electrical and Computer Engineering, Virginia Tech. He received his BS degree from Shandong University and MS degree from Tsinghua University in 2001 and 2004, respectively. Since Fall 2006, he started his graduate study at Virginia Tech under the supervision of Dr. Yue Wang and has been a graduate research assistant in the Computational Bioinformatics and Bio-imaging Laboratory of the department. His current research focuses on machine learning and its applications to high-throughput genomic and gene-expression data.

Honors and Awards

The William A. Blackwell Award by ECE department of Virginia Tech for the best research contribution, 2011.
The Best Student Paper at IEEE International Conference on Bioinformatics & Biomedicine, Washington, DC, 2009.
The First Prize of Shandong Province in **National Mathematical Modeling Contest,** P.R. China, 2001.

Publications

Journal Papers (7 papers produced during PhD study at Virginia Tech)

1. Li Chen, **Guoqiang Yu**, Carl D. Langefeld, David J. Miller, Richard T. Guy, Xiguo Yuan, David Herrington and Yue Wang, "Comparative analysis of methods for detecting interacting loci", ***BMC Genomics***, doi:10.1186/1471-2164-12-344**,** 2011

2. **Guoqiang Yu***, Bai Zhang*, G. Steven Bova, Jianfeng Xu, Ie-Ming Shih, Yue Wang, "BACOM: In silico detection of genomic deletion types and correction of normal cell contamination in copy number data", ***Bioinformatics***, doi: 10.1093/bioinformatics/btr183, 2011 (* joint first author)

3. **Guoqiang Yu**, Huai Li, Sook Ha, Ie-Ming Shih, Robert Clarke，Eric P. Hoffman, Subha Madhavan, Jianhua Xuan, Yue Wang, "PUGSVM: A caBIG^{TM} analytical tool for multiclass gene selection and predictive classification", ***Bioinformatics***, doi: 10.1093/bioinformatics/btq721, 2010

4. **Guoqiang Yu**, Yuanjian Feng, David J. Miller, Jianhua Xuan, Eric P. Hoffman, Robert Clarke, Ben Davidson, Ie-Ming Shih, and Yue Wang, "Matched gene selection and committee classifier for molecular classification of heterogeneous diseases," ***Journal of Machine Learning Research,*** vol.11, pp.2141-2167, 2010.

5. Yuanjian Feng, **Guoqiang Yu**, Tian-Li Wang, Ie-Ming Shih, and Yue Wang, "Analyzing DNA copy number changes using fused margin regression," ***Intl J of Functional Informatics and Personalized Medicine***, vol.3, no. 1, pp.3-15, 2010.

6. David J. Miller, Yanxin Zhang, **Guoqiang Yu**, Yongmei Liu, Li Chen, Carl D. Langefeld, David Herrington, and Yue Wang, "An Algorithm for Learning Maximum Entropy Probability Models of Disease Risk That Efficiently Searches and Sparingly Encodes Multilocus Genomic Interactions," ***Bioinformatics***, vol. 25, no. 19, pp. 2478-2485, 2009.

7. Wennuan Liu, Sari Laitinen, Sofia Khan, Mauno Vihinen, Jeanne Kowalski, **Guoqiang Yu**, Li Chen, Srinivasan Yegnasubramanian, Jun Luo, Yue Wang, Jianfeng Xu, William B. Isaacs, Tapio Visakorpi, and G. Steven Bova, "Copy Number Analysis Indicates Monoclonal Origin of Lethal Metastatic Prostate Cancer," ***Nature Medicine,*** vol. 15, no. 5, pp. 559-565, 2009.

8. Shiliang Sun, Changshui Zhang, **Guoqiang Yu**, "A Bayesian Network Approach to Traffic Flow Forecasting,", ***IEEE Transactions on Intelligent Transportation Systems***, vol. 7, no. 1, pp. 124-132, 2006.

9. Jianwei Gu, Li Zhang, **Guoqiang Yu**, Yuxiang Xing, Zhiqiang Chen, "X-ray CT Metal Artifacts Reduction through Curvature Based Sinogram Inpainting", ***Journal of X-Ray Science and Technology***, vol. 14, no. 2, pp.73-82, 2006.

Conference Papers (4 papers produced during PhD study at Virginia Tech)

1. **Guoqiang Yu**, Bai Zhang, Jianfeng Xu, Ie-Ming Shih, and Yue Wang, "Accurate Estimation of Genomic Deletions and Normal Cell Contamination by Bayesian Analysis of Mixtures", *IEEE Intl Conf. on Bioinformatics & Biomedicine*, Washington D.C., USA, Nov. 2009

2. Yuanjian Feng, **Guoqiang Yu**, Tian-Li Wang, Ie-Ming Shih, and Yue Wang, "Analyzing DNA Copy Number Changes Using Fused Margin Regression", *IEEE Intl Conf. on Bioinformatics & Biomedicine*, Washington D.C., USA, Nov. 2009.

3. Li Chen, **Guoqiang Yu**, David Miller, Lei Song, Carl Langefeld, David Herrington, Yongmei Liu, and Yue Wang, "A Ground Truth Based Comparative Study on Detecting Epistatic SNPs", *The 2009 Applications of Machine Learning in Bioinformatics Workshop*, Washington D.C., USA, Nov. 2009.

4. **Guoqiang Yu**, David Herrington, Carl Langefeld, and Yue Wang, "Detection of Complex Interactions of Multiple SNPs", *IEEE Intl Workshop on Machine Learning for Signal Processing*, Cancún, Mexico, October 16-19, 2008.

5. **Guoqiang Yu**, Jin Zhang, Zhiqiang Chen, Li Zhang, Yuxiang Xing, "Variational Segmentation of X-Ray Image with Overlapped Objects"**,** *SPIE Symposium on Electronic Imaging (**SPIE EI2006**) as an oral report,* January 15–19, 2006, San Jose, California, USA

6. **Guoqiang Yu**, Liang Li, Jianwei Gu, Li Zhang, "Total Variation Based Iterative Image Reconstruction"**,** I*CCV2005 workshop: Computer Vision for Biomedical Image applications* (***CVBIA***)*: Current Techniques and Future Trends,* October 15-21, Beijing, China

7. Jianwei Gu, Li Zhang, **Guoqiang Yu**, Zhiqiang Chen, Yuxiang Xing, "Metal Artifacts Reduction in CT Images through Euler's Elastica and Curvature Based Sinogram Inpainting"**,** *SPIE Symposium on Medical Imaging **(SPIE MI2006)***, February 11–16, 2006, San Diego, California USA

8. **Guoqiang Yu**, Li Zhang, Jin Zhang, "Combining Iterative Inverse Filter with Shock Filter for Baggage Inspection Image Deblurring," *Asian Conference on Computer Vision* (***ACCV2006***) as an oral report, January 13-16, 2006, Hyderabad, India

9. Shiliang Sun, **Guoqiang Yu**, Changshui Zhang, "Short-Term Traffic Flow Forecasting Using Sampling Markov Chain Method with Incomplete Data", *IEEE Intelligent Vehicles Symposium (**IV2004**)*, June 14-17, 2004, Parma, Italy.

10. Changshui Zhang, Shiliang Sun, **Guoqiang Yu**. "Short-Term Traffic Flow Forecasting Using Expanded Bayesian Network for Incomplete Data", *International Symposium on Neural Networks (**ISNN2004**)*, August 19-21, 2004, Dalian, China

11. **Guoqiang Yu**, Changshui Zhang, "Switching ARIMA Model Based Forecasting for Traffic Flow"**,** published on *International Conference on Acoustics, Speech, and Signal Processing (**ICASSP2004**)*, May 17-21, 2004, Canada

12. Shiliang Sun, Changshui Zhang, **Guoqiang Yu**, Naijiang Lu and Fei Xiao, "Bayesian Network Methods for Traffic Flow Forecasting with Incomplete Data", *the 15th European Conference on Machine Learning (**ECML 2004**)*, September 20-24, 2004, Pisa, Italy

13. Changshui Zhang, Shiliang Sun, **Guoqiang Yu**. "A Bayesian Network Approach to Time Series Forecasting of Short-Term Traffic Flows"**,** *the 7th IEEE International Conference on Intelligent Transportation Systems (**ITSC2004**)*, October 3-6, 2004, Washington, D.C, USA

14. Jianming Hu, Jingyan Song, **Guoqiang Yu**, Yi Zhang. "A Novel Networked Traffic Flow Forecasting Method Based on Markov Chain Model"**,** *IEEE International Conference on Systems, Mans & Cybernetic (**ICSMC2003**)*, October 5-8, 2003, Washington D.C, USA

15. **Guoqiang Yu**, Jianming Hu, Changshui Zhang, Like Zhuang, Jingyan Song, "Short-term Traffic Flow Forecasting Based on Markov Chain Model"**,** *IEEE Intelligent Vehicles Symposium (**IV2003**)*, June 9-11, 2003, Columbus, OH, USA

Machine Learning to Interrogate High-throughput Genomic Data: Theory and Applications

Guoqiang Yu

Abstract

The missing heritability in genome-wide association studies (GWAS) is an intriguing open scientific problem which has attracted great recent interest. The interaction effects among risk factors, both genetic and environmental, are hypothesized to be one of the main missing heritability sources. Moreover, detection of multilocus interaction effect may also have great implications for revealing disease/biological mechanisms, for accurate risk prediction, personalized clinical management, and targeted drug design. However, current analysis of GWAS largely ignores interaction effects, partly due to the lack of tools that meet the statistical and computational challenges posed by taking into account interaction effects. Here, we propose a novel statistically-based framework (Significant Conditional Association) for systematically exploring, assessing significance, and detecting interaction effect. Further, our SCA work has also revealed new theoretical results and insights on interaction detection, as well as theoretical performance bounds. Using *in silico* data, we show that the new approach has detection power significantly better than that of peer methods, while controlling the running time within a permissible range. More importantly, we applied our methods on several real data sets, confirming well-validated interactions with more convincing evidence (generating smaller p-values and requiring fewer samples) than those obtained through conventional methods, eliminating inconsistent results in the original reports, and observing novel discoveries that are

otherwise undetectable. The proposed methods provide a useful tool to mine new knowledge from existing GWAS and generate new hypotheses for further research.

Microarray gene expression studies provide new opportunities for the molecular characterization of heterogeneous diseases. Multiclass gene selection is an imperative task for identifying phenotype-associated mechanistic genes and achieving accurate diagnostic classification. Most existing multiclass gene selection methods heavily rely on the direct extension of two-class gene selection methods. However, simple extensions of binary discriminant analysis to multiclass gene selection are suboptimal and not well-matched to the unique characteristics of the multi-category classification problem. We report a simpler and yet more accurate strategy than previous works for multicategory classification of heterogeneous diseases. Our method selects the union of one-versus-everyone phenotypic up-regulated genes (OVEPUGs) and matches this gene selection with a one-versus-rest support vector machine. Our approach provides even-handed gene resources for discriminating both neighboring and well-separated classes, and intends to assure the statistical reproducibility and biological plausibility of the selected genes. We evaluated the fold changes of OVEPUGs and found that only a small number of high-ranked genes were required to achieve superior accuracy for multicategory classification. We tested the proposed OVEPUG method on six real microarray gene expression data sets (five public benchmarks and one in-house data set) and two simulation data sets, observing significantly improved performance with lower error rates, fewer marker genes, and higher performance sustainability, as compared to several widely-adopted gene selection and classification methods.

Table of Contents

List of Figures

List of Tables

I. Introduction

The advent of microarray technology (Schena, et al., 1995) and the completion of the International Hapmap Project (Thorisson, et al., 2005) ushered in the era of agnostic, very high-dimensional data measurement for the complex disease. While the microarray technique is a great advancement on the measurement of potential disease markers, it poses tremendous challenges, requiring new statistical and computational methods to extract useful information from the huge volume of microarray data. This dissertation focuses on several key challenges and tasks in analyzing high-throughput genomic and gene-expression data, which include detection of gene-gene and/or gene-environment interaction effects in determining complex diseases, and gene selection for multicategory disease prediction.

I.1 Motivation

The past several years have witnessed an avalanche of genome-wide association studies (GWAS) accelerated by the advent of affordable and accurate SNP chip technology (Sachidanandam, et al., 2001; Syvanen, 2001), which can simultaneously assay hundreds of thousands of ubiquitous and discriminative genetic markers --- Single Nucleotide Polymorphisms (SNPs). GWAS has been quite successful with over 400 genomic regions identified reproducibly predisposing to more than 70 common diseases or complex traits, including cancers, cardiovascular disease, diabetes, autism and height, etc. (Hindorff, et al., 2009; McCarthy, et al., 2008), and the discovering process is still ongoing. However, concerns and discomfort is escalating among researchers who are realizing that the variants discovered until now explain only a small fraction of genetic

contribution to most diseases or traits (Maher, 2008; Manolio, et al., 2009). For even heavily studied diseases the variants found to date typically explain less than 20% of the heritable variance in disease risk (Goldstein, 2009). In addition, even for the well-established discoveries emerging from GWAS, it is proving hard to interpret the biological functions and implications as most of the detected SNPs reside in the introns or nowhere near any gene (Hindorff, et al., 2009; McCarthy, et al., 2008).

One of the main contributors to the missing heritability is widely considered to be the interaction effects among risk factors, such as the gene-gene interaction and gene-environment interaction. While the hypothesis of multi-locus effects on complex diseases has long been suggested and increasing empirical evidence from both model organisms and human studies suggests that the interaction of multiple loci contributes broadly to complex diseases, most of the current genome-wide association studies detect disease-susceptibility SNPs based only on the single locus evaluation, perhaps largely due to the lack of powerful and accessible tools to analyze multi-locus effects. The complexity and scale of interaction in the context of genome-wide association studies poses daunting challenges on the design of detection criteria, the evaluation of significance, and the development of computationally practical yet fruitful interaction searching schemes.

The scenario may exist wherein none of the interacting SNPs shows significant association individually with the condition in question, but where the joint evaluation with the interaction effect taken into account could provide sufficient support to the association. One may argue that the consideration of joint effects will incur a huge number of multiple tests and thus dramatically

tighten the experimental-wise significance threshold. Our theoretical analysis in the following clearly indicates that, under modest conditions that relate to the effect size and the allele frequency, incorporating interaction effects will surpass the cost introduced by the multiple tests and improve the detection power of informative SNPs.

Besides capturing missing heritability, the identification of interaction among multiple SNPs *per se* is very important and useful with significant consequences, such as inferring pathways, illuminating the disease mechanism, improving the prediction power for clinical use, and developing targeted drugs. This task (identifying the interaction effects among significant SNPs) and the task described in the above paragraph (detecting disease-susceptibility or informative SNPs by incorporating the interaction effects) are two related yet distinct tasks, which are often vaguely expressed and muddled together in practice, resulting in suboptimal solutions. If a group of multiple SNPs without significant marginal effects is found to be significantly associated with disease through joint analysis with the interaction effects incorporated, then these SNPs are interacting with each other by definition. However, the identification of interaction among multiple individually significant SNPs needs more careful design and cannot replace the task of improving detection power of informative SNPs, because it usually removes the marginal effects and has less sensitivity to detect informative SNPs.

The incorporation of interaction effects requires intensive computing power and exhaustive search is infeasible for even two-order interactions on typical GWAS datasets containing hundreds of thousands SNPs and several thousands of subjects. An efficient and powerful heuristic searching scheme is therefore warranted and indispensable. To be an effective

searching algorithm, it should have the ability to reduce significantly the running time with a small (controlled) degree of deterioration of performance. Usually, the heuristic searching algorithm provides users some parameters to control the tradeoff between running time and performance. It would be desirable for the control parameter to be *directly* linked to the performance; thus, the users will know what they can expect to achieve from the search, without struggling to estimate how likely it is that they have missed real important findings because of the heuristic nature of the search.

Despite the realities concerning the wide variety of possible multi-locus relationships, previous efforts to study genetic interactions have typically relied on linear models and multiplicative interaction terms, which are largely driven by mathematical convenience and the familiarity of logistic models of binary traits (Agresti, 2002). However, these methods for primarily detecting independent main effects are ill-suited to discover nonlinear interaction effects, and have only produced limited success, due to high numbers of false positives and false negatives. Several computational and machine learning approaches have recently been proposed to decipher the interaction puzzle embedded in GWAS studies, including Multifactor Dimensionality Reduction (MDR) (Ritchie, et al., 2001), Logic Tree (Ruczinski, et al., 2003), Full Interaction Model (FIM) (Marchini, et al., 2005), Bayesian Epistasis Association Mapping (BEAM) (Zhang and Liu, 2007), Penalized Logistic Regression (PLR) (Park and Hastie, 2008) and SNP Harvester (Yang, et al., 2009). These existing methods have achieved only minor acceptance in GWAS studies due to various reasons, one of which is the ambiguity between incorporating and identifying the interaction effects, resulting in high false positive rates for both marker discovery and interaction effects identification. Secondly, for most of the methods, a definite interaction model is required,

such as PLR and SNP Harvester, whereas interactions in biology are complex, with no particular, single model achieving wide consensus. Thirdly, some methods, like MDR and Logic Tree, applied cross validation and permutation testing to evaluate markers and their significance. This is intrinsically computation-intensive and its application is prohibited on large datasets. Fourthly, even if some methods provide schemes to accelerate the search, they typically involve greedy search or heuristic search, whose success highly relies on the presence of significant marginal effects; yet, we know that complex and nonlinear interactions often show little or no marginal effects. Lastly, and probably most critical for the existing methods less accepted by terminal users is that, although users have been given the freedom to control the accuracy and running time, the parametric dependence on both execution time and detection performance is complicated. Typically, it is completely unknown, for these methods, what is the likelihood for a real interaction being missed due to insufficient running time under the given setting.

The rapid development of gene expression microarrays provides an opportunity to take a genome-wide approach for disease diagnosis, prognosis, and prediction of therapeutic responsiveness (Clarke, et al., 2008; Wang, et al., 2008). When the molecular signature is analyzed with pattern recognition algorithms, new classes of disease are identified and new insights into disease mechanisms and diagnostic or therapeutic targets emerge (Clarke, et al., 2008). For example, many studies demonstrate that global gene expression profiling of human tumors can provide molecular classifications that reveal distinct tumor subtypes not evident by traditional histopathological methods (Golub, et al., 1999; Ramaswamy, et al., 2001; Shedden, et al., 2003; Wang, et al., 2006).

While molecular classification falls neatly within supervised pattern recognition, high gene dimensionality and paucity of microarray samples pose challenges for, and inspire novel developments in classifier design and gene selection methodologies (Wang, et al., 2008). For multicategory classification using gene expression data, various classifiers have been proposed and have achieved promising performance, including k-Nearest Neighbor Rule (kNN) (Golub, et al., 1999), artificial neural networks (Wang, et al., 2006), Support Vector Machine (SVM) (Ramaswamy, et al., 2001), Naïve Bayes Classifier (NBC) (Liu, et al., 2002), Weighted Votes (Tibshirani, et al., 2002), and Linear Regression (Fort and Lambert-Lacroix, 2005). Many comparative studies show that SVM based classifiers outperform other methods on most bench-mark microarray datasets (Li, et al., 2004; Statnikov, et al., 2005).

An integral part of classifier design is gene selection, which can improve both classification accuracy and diagnostic economy (Liu, et al., 2002; Shi, et al., 2008; Wang, et al., 2008). Many microarray-based studies suggest that, irrespective of the classification method, gene selection is vital for achieving good generalization performance (Statnikov, et al., 2005). For multicategory classification using gene expression data, the criterion function for gene selection should possess high sensitivity and specificity, well match the specific classifiers used, and identify gene markers that are both statistically reproducible and biologically plausible (Shi, et al., 2008; Wang, et al., 2008). There are limitations associated with existing gene selection methods (Li, et al., 2004; Statnikov, et al., 2005). While wrapper methods consider joint discrimination power of a gene subset, complex classifiers used in wrapper algorithms for small sample size may overfit, producing non-reproducible gene subsets (Li, et al., 2004; Shi, et al., 2008). Moreover,

discernment of the (biologically plausible) gene interactions retained by wrapper methods is often difficult due to the black-box nature of most classifiers (Shedden, et al., 2003).

Conversely, most filtering methods for multicategory classification are straightforward extensions of binary discriminant analysis. These methods are devised without well-matching to the classifier that is used, which typically leads to suboptimal classification performance (Statnikov, et al., 2005). Popular multicategory filtering methods (which are extensions of two-class methods) include Signal-to-Noise Ratio (SNR) (Dudoit, et al., 2002; Golub, et al., 1999), Student's t-statistics (Dudoit, et al., 2002; Liu, et al., 2002), the ratio of Between-groups to Within-groups sum of squares (BW) (Dudoit, et al., 2002), and SVM based Recursive Feature Elimination (RFE) (Li and Yang, 2005; Ramaswamy, et al., 2001; Zhou and Tuck, 2007). However, as pointed out by Loog et al. in proposing their weighted Fisher criterion (wFC) (Loog, et al., 2001), simple extensions of binary discriminant analysis to multicategory gene selection are suboptimal because they overemphasize large between-class distances, i.e. these methods choose gene subsets that preserve the distances of (already) well-separated classes, without reducing (and possibly with increase in) the large overlap between neighboring classes. This observation and the application of wFC to multicategory classification are further evaluated experimentally by Wang et al. and Xuan et al. (Wang, et al., 2006; Xuan, et al., 2007).

I.2 Objectives and Statement of Problems

I.2.1 Analyzing Interaction Effects in Complex Diseases

In this dissertation we propose a systematic approach, namely, Significant Conditional Association (SCA), to tackle problems in GWAS caused by interaction effects. The consideration of the interaction effects introduces two significant challenges, both statistically and computationally. When interactions are referred, three are usually two objectives (implicitly) in mind. The first objective is to incorporate the interaction effect in the process of marker discovery, to improve the detection power of disease-associated markers. The second objective is to identify the interaction effect among the established disease factors, with the aim of shedding light on the biological mechanism. It is challenging to design statistical criteria to take advantage of interaction effects, since the statistical criterion needs to distinguish the two objectives, have high sensitivity and high specificity, and should afford accurate assessment of statistical significance. The computation is another daunting issue and an efficient and user-friendly heuristic searching scheme is in great demand.

Accordingly, to analyze interaction effects in complex diseases we would like to seek the solutions to the following tasks:

1. Systematically model and statistically evaluate interaction effects to find novel disease-risk factors.
2. Systematically model and statistically evaluate interaction effects among established disease-risk factors.
3. Develop an efficient and user-friendly heuristic search strategy.

I.2.2 Gene Selection for Multi-category Disease Prediction

Multiclass gene selection is an imperative task for identifying phenotype-associated mechanistic genes and achieving accurate diagnostic classifications. Statistical reproducibility and biological plausibility are two key requirements for a good gene selection approach, due to the high dimensionality and the small sample size in microarray experiments. The multi-category classification problem also has its unique characteristics. Unlike the binary case, there are multiple pairs of classes and the distance can be quite different for different pairs of classes. Accordingly, for the multi-category gene selection problem, we set the following research objective:

1. Develop a multi-category gene selection method that has accurate disease prediction, is statistically reproducible, is biologically plausible, and is robust to unbalanced class distributions.

I.3 Organization of the Dissertation

In this chapter, we have briefly introduced the background knowledge. After describing our motivations, two problems were identified and the associated challenges were discussed. The research objectives were also explicitly listed.

In Chapter II, we will describe with full details the proposed approach to the multi-locus interaction effects of complex diseases. We will also present rigorous mathematical proofs justifying the proposed approach and experimental results on both simulation and real datasets. Following the similar pattern, Chapter III focuses on multi-class gene selection for gene

expression data. A thesis summary and the proposed future research work are discussed in Chapter IV.

II. SCA to Identify Complex Interactions in GWAS

To address the challenges and opportunities inherent in detection of complex multi-locus interaction effects associated with complex phenotypes in GWAS data, we propose a systematic approach, *Significant Condition Association* (SCA), to characterize and take advantages of the interaction effects, supported by rigorous analysis and extensive experiments. This chapter includes two sections. The first section discusses the methods and theory. The second section demonstrates the experimental results.

II.1 Methods and Theory

Recalling that in practice there are two different tasks pertaining to the interaction effects in GWAS studies, we have proposed tailored approaches to each of the tasks. The global picture of the proposed SCA framework is demonstrated in Figure 1 for the analysis of up to second order interactions. The algorithm takes the genotypes as the input and consists of three key modules. Module I detects informative SNPs with interaction effect incorporated, corresponding to the objective 1 mentioned earlier. Module II identifies interactions among significant SNPs, which corresponds to objective 2. And Module III performs the heuristic search to generate the k-tuple candidates, dealing with the daunting computational challenges due to the exponentially growing number of SNP combinations. The algorithm reduces to Fisher's exact test when we restrict ourselves to individual-SNP analysis only, of which both the module II and the module III are

irrelevant and are skipped. Otherwise, all three modules are repeatedly applied for the analysis at each order. Take the two-way interaction analysis as an example. If there is at least one insignificant SNP in the pair of SNPs, Module I is called to assess the significance of the pair. Module II is applied to evaluate the pairs for which each SNP is significant. The construction of the candidate pair from the pool of all insignificant SNPs needs to resort to heuristic search, that is, Module III, considering that the size of the pool is usually very large.

Although Figure 1.a only shows the analysis of up to two-way interactions, higher order analysis is a straightforward extension and is actually implemented in our software. Generally speaking, for the k^{th} order analysis, as long as there is one SNP not marked yet as significant in the k-tuple, Module I will be invoked, otherwise, Module II will be applied. Module III will be used in the case that the candidate k-tuple to Module I contains at least two insignificant SNPs. Please be noted that a heuristic search strategy is not generally necessary for generating the candidate SNP subset for Module II, since the number of significant SNPs is much smaller than the total size of SNPs assayed in the study and exhaustive search over this subset is typically feasible.

In Subsection II.1.1 we discuss the algorithm and the theory associated with the first task of incorporating the interaction effect. The detection power of the disease-susceptibility markers is expected to be improved by incorporating the gene-gene or gene-environment interaction effects. Then, we discuss in Subsection II.1.2 the second task, that is, to determine whether the interaction effect itself exists among multiple marginally significant disease factors. As mentioned before, the mammoth number of SNPs in GWAS and the large sample size make the computation a daunting mission. Exhaustive search is simply infeasible for detecting multi-locus interactions and heuristic search is a necessity. The computation is so challenging and important

a subject that it deserves a full discussion in a separate subsection, as we do so in Subsection II.1.3.

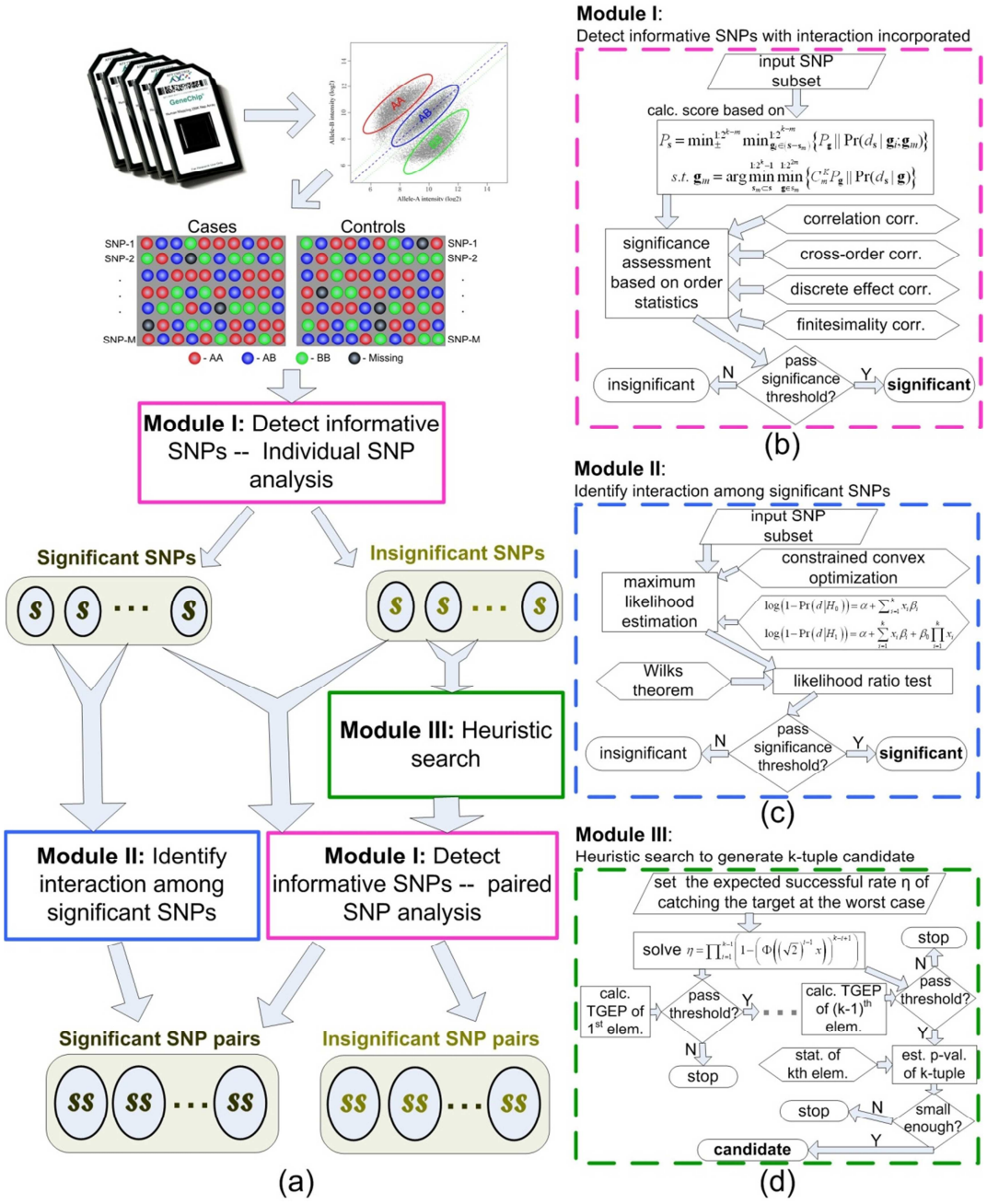

Figure 1: Overview of the SCA framework. The goal of the algorithm is to embrace the interaction effect into the analysis of GWAS data, either to improve the detection rate of informative markers or to identify the interaction effect among significant/known markers.

II.1.1 Detecting Informative SNPs by Incorporating Interaction Effect

The proposed approach in this subsection aims at the problem of incorporating interaction effects to improve marker detection power. We first demonstrate the principle of incorporating interaction effect in the discovery process by dissecting a set of simple illustrative examples. Accordingly, we transform this principle into a mathematical testing statistics and the connection is established between the mathematical formulae and the principle. The implication of the proposed statistics is further discussed with respect to the flexible dominant/recessive genetic model, high specificity and high sensitivity, and the reduced search space. The hyper-geometric probability model serves as the basic building block for our approach. The importance of adopting this probability model is mainly due to its exactness. However, the main critique to use of the hyper-geometric model is its high computational demand. We develop a fast approximate algorithm to speed the computation while controlling the error such that more than 95% accuracy will be guaranteed. All these topics are discussed in subsubsections II.1.1.A~D.

The proposed statistic readily suggests a way to utilize the existing discoveries to further reduce the computational burden. The re-utilization strategy is discussed in subsubsection II.1.1.E. The assessment of significance is a critical part of a statistical approach of discovery. Exact assessment is almost impossible largely because of the existing correlation among the component random variables. We present a computationally-feasible approximation to the p-value of the proposed statistic in subsubsection II.1.1.F. The conservativeness of the approximation is discussed in subsubsection II.1.1.G and our theoretical analysis shows that the approximation is in fact very good within the context of GWAS studies. Finally, the issue of missing values is discussed in subsubsection II.1.1.H.

II.1.1.A Principle of incorporating interaction effect in marker discovery

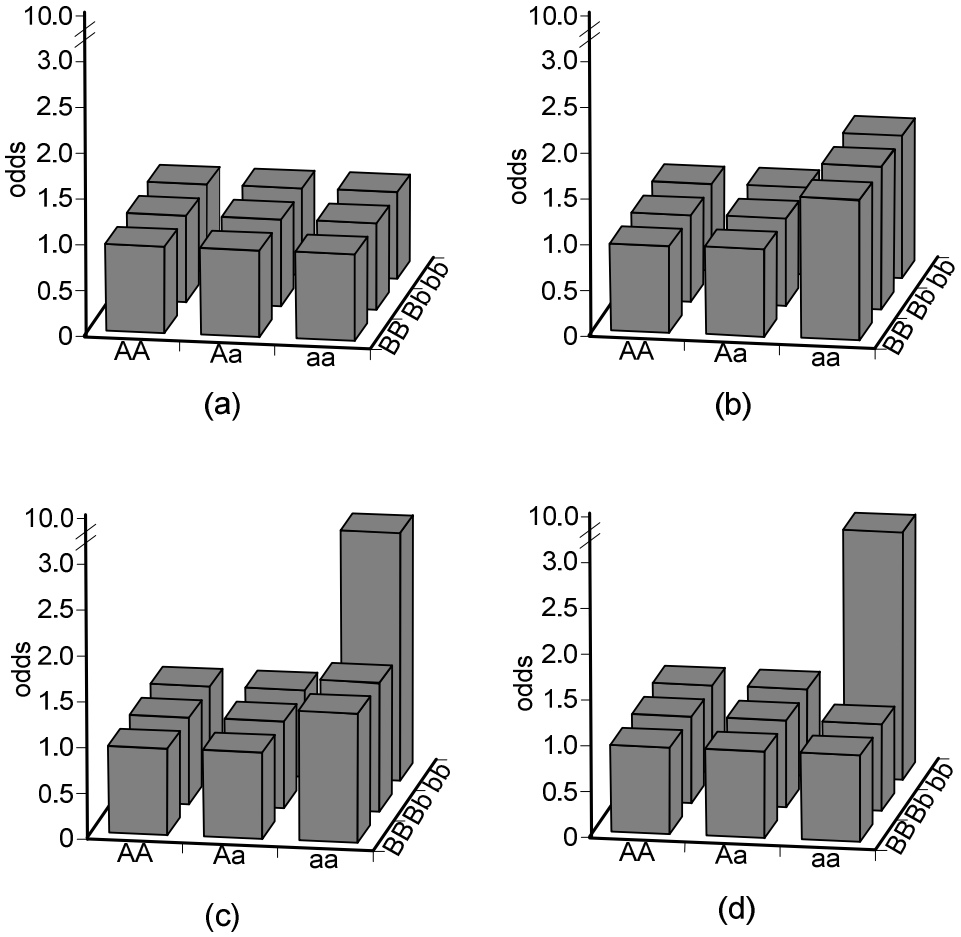

Figure 2: Four typical scenarios involving a pair of two SNPs. The two SNPs do not interact with each other in the first two scenarios and the two SNPs show interaction effects in the last two scenarios.

A set Φ of multi-locus SNPs is considered to be interactingly conferring disease status if the set as a whole provides more information about the likelihood of the phenotype than the maximum information provided by any of its lower-order subsets. Two points deserve to be emphasized here. First, the SNP set has information on the likelihood of the phenotype; secondly, it has additional information, beyond that contained in any of its lower order subsets. Thus, a coupled criterion is proposed to evaluate the SNP subset as follows: (1) the distribution of the phenotype

changes at least at one of the genotypes, and (2) each member of Φ has *intrinsic influence* on the phenotype.

Let us see how the proposed coupled criteria work in the four illustrative examples as shown in Figure 2. By looking at these four subfigures, we may have the rough impression that (a) and (b) should not be considered as jointly informative, while (c) and (d) are jointly informative.

The genotypes specified by SNP A and SNP B in subfigure (a) do not change the disease risk at all. So SNP-A and SNP-B fail criterion number 1. The SNP subset is not informative, without even considering whether they interactingly confer disease status.

For the second scenario in subfigure (b), the genotypes specified by SNP A and SNP-B do change the disease risk and criterion number 1 is passed. However, we notice that given the status of SNP-A, the genotypes specified by SNP-B do not change the disease risk, that is, all the odds ratios are the same along the right column. So, the SNP subset in subfigure (b) fails criterion number 2 and should not be considered as interactingly conferring the disease status.

For the third scenario, as in the subfigures (c), we see that the genotypes in the right column are associated with different disease risks from other genotypes. These two SNPs are compatible with criterion number 1. We further find that, given the status that SNP A = aa, the status of SNP B has influence on the disease risk, i.e., the genotype {aabb} has different disease risk than under the genotypes {aaBB} and {aaBb}. The SNP subset passes both two criteria and should be considered as interactingly conferring disease status.

The fourth scenario in the subfigures (d) is quite similar to the third scenario. It is easy to verify that the SNP subset passes both two criteria and thus should be considered as jointly informative. The difference is that SNP-A in subfigure (c) is informative itself, while no SNP in subfigure (d) shows significant marginal effect.

II.1.1.B Statistical measure of a SNP subset with interaction taken into account

Supposing there are *K* SNPs in total and *N* samples, with *D* samples labeled as cases. Motivated by the aforementioned coupled criteria, we propose the following measure to evaluate a set Φ of *k* SNPs with the interaction taken into account:

$$P_{\Phi} = \min_{\pm}^{1:2^{k-m}} \min_{g_n \in (\Phi \cap \overline{\mathbf{s}_m})}^{1:2^{k-m}} \left\{ P_g \,||\, \Pr(d_{\mathbf{s}} \,|\, g_n; g_m) \right\} \tag{1}$$

$$\textit{subject to:}\quad g_m = \arg\min_{g} {}_{\mathbf{s}_m \subset \Phi}^{1:2^k-1} \min_{g \in \mathbf{s}_m}^{1:2^{2m}} \left\{ \binom{K}{m} \times T \times P_g \,||\, \Pr(d_{\mathbf{s}} \,|\, g) \right\}$$

where,

$$P_g = 2\min\left(P_L, P_R\right)$$

$$P_L = \sum_{t=\max(0,\, j+r-M)}^{d_{\mathbf{s}}} p(t; M, j, r)$$

$$P_R = \sum_{t=d_{\mathbf{s}}}^{\min(j,r)} p(t; M, j, r)$$

$$\Pr\left(d_{\mathbf{s}} \,|\, g_n; g_m\right) = p\left(d_{\mathbf{s}}; M, j, r\right) = \frac{\binom{j}{d_{\mathbf{s}}}\binom{M-j}{r-d_{\mathbf{s}}}}{\binom{M}{r}}$$

$$\Pr\left(d_{\mathbf{s}} \mid g\right) = p\left(d_{\mathbf{s}}; N, D, r\right) = \frac{\binom{D}{d_{\mathbf{s}}}\binom{N-D}{r-d_{\mathbf{s}}}}{\binom{N}{r}}.$$

Please be reminded that this is an important yet quite complicated equation, which will require the next 13 pages to fully explore its potentials. The equation can be understood in the following concise way. Small P_{Φ} implies 'having information' corresponding to criterion #1, while the conditional genotype g_m assures 'having intrinsic information' corresponding to criterion #2.

Intuitively speaking, g_m is a genotype, serving as the condition for the optimization problem $\min_{\pm}^{1:2^{k-m}} \min_{g_n \in (\Phi \cap \overline{\mathbf{s}_m})}^{1:2^{k-m}} \left\{P_g \,\|\, \Pr(d_{\mathbf{s}} \mid g_n; g_m)\right\}$, which searches for the most informative genotype. The degree of being informative is measured by the p-value P_g that is calculated based on the hypergeometric probability model $\Pr(d_{\mathbf{s}} \mid g_n; g_m)$. Take a set of examples shown in Figure 3. g_m is null for both examples (a) and (d). g_m is determined to be the genotype {aa} for both examples (b) and (c). Then, for examples (a) and (d), the optimization problem is to find the most significant one among 16 genotypes. Here, there are four dominant/recessive codes for two SNPs and there are four genotypes under each code. For each genotype, the hypergeometric probability model, which requires a binary variable, is constructed by considering the genotype in question as a group and all the other genotype as another group. Similarly, for examples (b) and (c), the optimization problem is to find the most significant one among 4 genotypes under the condition of the genotype {aa}.

Specifically, T is the number of equivalent independent multiple tests when searching for the best genotype specified by the SNP subset $\mathbf{s}_m \subset \Phi$. A quite loose upper bound is $2^m\left(2^m-1\right)$, with a more accurate procedure to calculate T to be described in equation (7). $\binom{K}{m}$ is the number of multiple tests coming from finding the best subset with m SNPs out of a total K SNPs. Genotype g_m specifies a population with M subjects, of which, j subjects are cases, $M-j$ subjects are controls, and r subjects carry genotype g_n. The probability that t out of the r subjects are cases is given by $p(t;M,j,r)$, which is exactly the hypergeometric distribution (Agresti, 2002). For the "observed" number of cases $d_{\mathbf{s}}$ in the r subjects specified by genotype g_n, an associated two-tailed p-value P_g conditioned on genotype g_m gives the probability of obtaining an observation at least as extreme as $d_{\mathbf{s}}$, assuming g_n has no intrinsic influence on disease status. Explicitly, associated with g_m (the genotype as the condition) and g_n (the genotype under the investigation), the parameter set (M,j,r), where M is the number of total subjects carrying g_m, j is the number of cases carrying g_m and r is the number of subjects carrying g_n, conveys the "message" about the null hypothesis $\Pr(d_{\mathbf{s}} \mid g_n; g_m)$. A significant deviation in $d_{\mathbf{s}}$ from such "specification" would produce a small P_g that subsequently indicates the intrinsic influence by g_n on disease status.

II.1.1.C Working examples on how to incorporate the interaction effect to improve power in marker discovery

Consider a GWA study of 5,000 subjects genotyped on 1,000,000 SNPs. Figure 3 shows how the algorithm works for four typical scenarios involving a pair of two SNPs. In all examples, the

MAF of both SNPs is 0.40, with the lower case indicating the minor allele. We assume that the two SNPs are statistically independent to each other and both satisfy Hardy-Weinberg equilibrium. As we have explained in subsection II.1.1.A, in the first two scenarios the two SNPs do NOT interactively influence the disease risk, while in the last two scenarios the two SNPs DO. The algorithm correctly evaluates them as statistically insignificant for the first two scenarios and statistically significant for the last two scenarios.

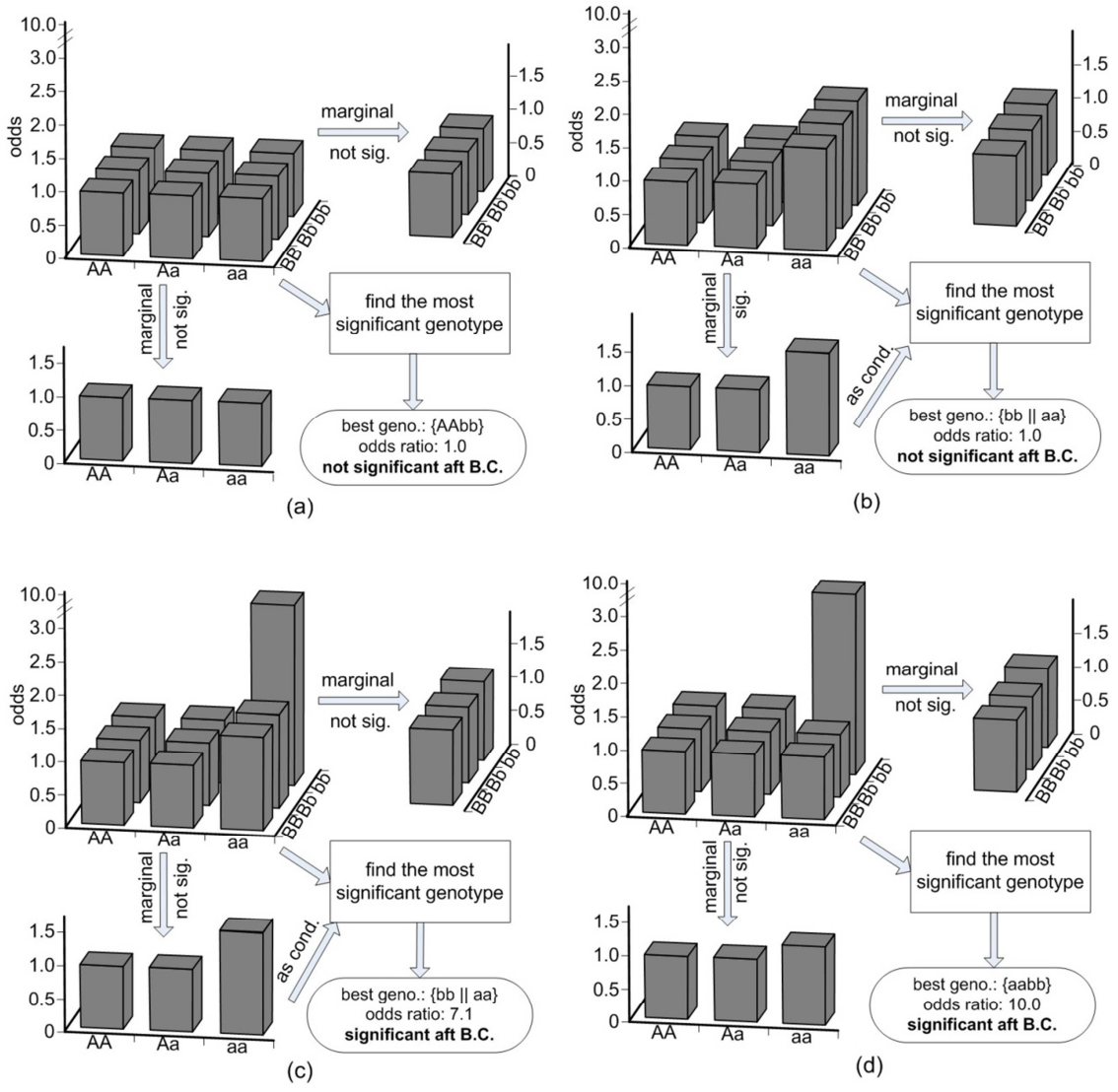

Figure 3: Illustrative examples on how the SCA algorithm incorporates interaction effects to improve power in marker discovery.

Scenario (Figure 3.a): neither of the SNPs is individually associated with the phenotype, nor interactively. All the genotypes have the same odds of 1.0. By evaluating the marginal effect, neither of the two SNPs is significant. Hence, there is no condition for finding the best genotype. Searching all the genotypes specified by the two SNPs, we find the genotype {AAbb} is the one with the smallest p-value, however, purely due to the random effect. And we know that the genotype {Aabb} has an expected odds ratio of 1.0; thus the two SNP are determined as 'not significant'.

Scenario (Figure 3.b): SNP-A is individually associated with the phenotype, but SNP-A and SNP-B are not interactively associated with the phenotype. The odds are 1.6 for genotypes {aaBB}, {aaBb} and {aabb} and 1.0 for all others. Looking at the marginal effect of SNP-A, we can see that the odds ratio for the genotype {aa} is 1.6 and the p-value is 1.07e-9. After the Bonferroni correction of 1,000,000 tests, SNP-A is significant. So, {aa} will be considered as a condition. And we know that SNP-B has an odds ratio of 1.0 conditioned on the genotype {aa}. Hence, SNP-A and SNP-B do not interactively influence the phenotype.

Scenario (Figure 3.c): SNP-A is individually associated with the phenotype and SNP-B is not, and they are interactively associated with the phenotype. The odds are 1.4 for genotypes {aaBB} and {aaBb}, 10.0 for the genotype {aabb}, and 1.0 for all others. For SNP-A, the odds ratio of {aa} is 1.74 with the corresponding p-value of 1.21e-12. SNP-A is significant after the Bonferroni correction. The genotype {aa} is chosen as a condition to search for the best

genotype specified by SNP-B. The genotype {bb || aa} has the smallest p-value of 4.37e-17. Here '|| aa' denotes the genotype {aa} is a condition. Noticing the number of the multiple tests is 1,000,000, SNP-B conditioned on SNP-A is significant after the Bonferroni correction. So, SNP-A and SNP-B are interactively associated with the phenotype.

Scenario (Figure 3.d): none of the SNPs is individually associated with the disease, but they are interactively associated with the phenotype. The odds are 10.0 for the genotype {aabb} and are 1.0 for all others. Projected on the margins, we obtain that the odds are 1.30 for both {aa} and {bb}. The two SNPs are not significant individually after the Bonferroni correction (the original p-value is 3.60e-4). So there is no condition for the search of the best genotype specified by the two SNPs. We find that genotype {aabb} is the one with the smallest p-value out of the 16 genotypes (4 genotypes per genetic model multiplied by 4 dominant-recessive genetic models). The p-value associated with {aabb} is 3.14e-20, which is 2.51e-7 after the Bonferroni correction and is significant, experimental-wisely.

If we do not take into account the interaction effect, in the example of (c) SNP-B will be missed, and in the example of (d) both SNP-A and SNP-B will be missed.

II.1.1.D More explanation of the proposed measure in equation (1)

The double-*min* operation in the **objection function** complies with **criterion #1**. It interrogates every possible genotype to hunt for the most significant genotype associated with the disease. This operation brings high sensitivity of marker discovery, since as long as one of the many possible genotypes influences disease susceptibility the SNP set will be detected as significant.

The double-*min* operation in the **constraint** takes care of **criterion #2**. It prevents the occurrence of the severe situation that a SNP subset gets a small p-value implying the association with the disease because its sub-subset is a real association, but withnot all the SNP in the subset *interactingly* contributing to the association. Otherwise, a real disease-associated multi-locus SNP group will contaminate its expanding SNP subset, resulting in a lot of false positives. Thus, the double-*min* operation in the constraint helps gain high specificity.

$\min_{\pm}^{1:2^{k-m}}$ allows the algorithm to flexibly fit the best dominant-recessive model. A SNP subset Φ with *k* members has 3^k genotypes and each genotype has on average $N/3^k$ samples, assuming the total number of samples is *N*. The detection power decreases significantly with reduced samples. Supported by the generally accepted concept that most of genetic diseases involve either dominant or recessive mutations (or both), the SNP status can be dichotomized and the average sample size per genotype increases from $N/3^k$ to $N/2^k$, which is a significant increase especially when *k* is large. Since each SNP could be either recessive or dominant, there are totally 2^k dichotomy models to recode the SNP subset Φ. The operation $\min_{\pm}^{1:2^{k-m}}$ is based on the condition that the genotype of *m* members in Φ is already specified as g_m, so there are only 2^{k-m} degrees of freedom left to search.

$\min_{g_n \in \left(\Phi \cap \overline{\mathbf{s}_m}\right)}^{1:2^{k-m}}$ allows us sensitively detect any unusual genotype that may imply the association with the disease, where $g_n \in \left(\Phi \cap \overline{\mathbf{s}_m}\right)$ means g_n is a genotype specified the SNPs that belong to Φ but do not belong to $\mathbf{s}_m$ (and there are totally 2^{k-m} such genotypes). Unlike the conventional methods such as chi-squared tests that sum the square of the deviation of each genotype, we use

the most significant deviation associated with any of the genotypes to represent the influence of $\left(\Phi \bigcap \overline{\mathbf{s}_m}\right)$ on the disease. When there is only one genotype significantly influencing the disease status, the "min" representation is more sensitive to detect the association than the "sum" representation, since the "sum" representation introduces many more degrees of freedom, which will make the p-value larger than that based on the multiple tests in the "min" representation. Although there is the possibility that more than one genotype influences the disease status, we expect that the chance will typically be small. Even in the event that it does happen, our "min" representation still has reasonable or better power to discover markers because we do not know *a priori* which genotypes take effect and the relative effect sizes.

The minimization $\min_{\mathbf{s}_m \subset \Phi}^{1:2^k-1} \min_{g \in \mathbf{s}_m}^{1:2^{2m}}$ in the constraint promises that, if P_Φ is small, all the members in Φ should have intrinsic contribution to the disease susceptibility, where g_m serves as a condition specified by the members in the proper SNP subset $\mathbf{s}_m \subset \Phi$. Here, the proper subset means that $\mathbf{s}_m$ is strictly contained in Φ and so necessarily excludes at least one member of Φ. There are totally $2^k - 1$ different sets of $\mathbf{s}_m$ with $m = 0,1,...,k-1$ denoting the size of $\mathbf{s}_m$. Let us see how the minimization in the constraint prevents assigning a significant value to the whole set Φ. For example, suppose $\Omega \subset \Phi$ is a true marker set associated with disease, but $\Phi \bigcap \overline{\Omega}$ is not. Then, under the condition of genotypes specified by Ω, no genotype specified by $\Phi \bigcap \overline{\Omega}$ will generate a small p-value since $\Phi \bigcap \overline{\Omega}$ is not associated with the disease. Accordingly, the final P_Φ will not be small. To select the most relevant genotype in the constraint, one needs to take the effects of multiple tests into account to make sure the comparison between different orders is fair. We use the Bonferroni correction to address this problem, that is, the effect of higher order

is penalized by the number of multiple tests. As noted above, $\mathbf{s}_m \subset \Phi$ is a proper subset of Φ, which implies that $\mathbf{s}_m$ can be the empty set ϕ with zero members in it. When $\mathbf{s}_m$ is empty, $P_{g\in\phi}$ in the constraint is undefined since no genotype is specified. Here, we define $P_{g\in\phi} = 0.05$ as default, which is consistent with P_g being calculated the other way and simplifies the further analysis that will be explained in the following. We provide the user the freedom to define $P_{g\in\phi}$ in their own way, such as $P_{g\in\phi} = 0.01$, as long as $P_{g\in\phi}$ is chosen as the significance threshold.

With the measure defined in equation (1), a mixture of multiple low-order interaction models is allowed to define complex associations that accommodate a wide variety of composite interactions, reflecting biological multimodality. Our mixture interaction model makes an explicit reflection of disease heterogeneity. Supposing there are Q risk modules for a certain disease, with each module specified by Λ_i interacting SNPs, $i = 1,...,Q$, then we can define the following two numbers:

$$\begin{aligned} N_T &= \sum_{i=1}^{Q} \Lambda_i \\ N_M &= \max_i^{1:Q} \{\Lambda_i\} \end{aligned} \qquad (2)$$

N_T is the total number of susceptibility SNPs to the disease, and N_M is the maximum order of interaction models. $N_T >> N_M$ for complex diseases since $\Lambda_i \geq 1$ and Q will be fairly large. A method trying to evaluate jointly all the susceptibility SNPs will search up to N_T to find all the markers, while our method only searches up to N_M, which will gain a huge margin on the computation considering $N_M << N_T$. Another benefit for a small number of searching order is

that it is much more robust to false positives because our method suffers at most N_M possible sources of errors whereas a joint evaluation approach suffers N_T possible sources of errors.

II.1.1.E Importance of the hyper-geometric probability model and its fast implementation

We choose the hyper-geometric distribution as the basic probability model, which turns out to be important for the whole approach's success not only because of its exactness but also because it facilitates an efficient heuristic searching algorithm with theoretical justification and a more accurate approximation to the final p-value based on the order statistics. The hyper-geometric distribution is an exact model in the sense that, no matter what the setting is, the probability calculated is always correct, while the approximation approach to the contingency table such as a Gaussian approximation will have a large error when the table is skewed or the cell contains too few samples. Another characteristic of the hyper-geometric distribution is that, given the parameters the minimum of possible p-values is determinate and strictly larger than 0 as shown in ***Proposition 1***, compared to a value that may approach to 0 from the Gaussian approximation. This seemingly trivial feature turns out to have important practical implications. To get a smaller p-value than the predefined experimental-wise significance threshold, parameters in the hyper-geometric distribution must satisfy a certain condition; hence some SNPs subsets or genotypes can be relieved from evaluation based on an easy check of this condition. In addition, the proposed measure in equation (1) involves some dependent multiple tests, which requires that the calculation of the final p-value be carefully designed. The characteristic of a finite p-value associated with the hyper-geometric distribution makes it possible to construct a good approximation to the final p-value.

Proposition 1: **Assume the random variable** v **follows the hyper-geometric distribution such that the probability** $\Pr(v=t)=\binom{j}{t}\binom{M-j}{r-t}\Big/\binom{M}{r}$**, where** M **,** j **and** r **are parameters. Denote** $\mathrm{H}(b;M,j,r)$ **as the p-value associated with the observed value** b**. Then, the minimum possible p-value under the hyper-geometric distribution is,** $\min_b\{\mathrm{H}(b;M,j,r)\}=\binom{\max(j,r)}{\min(j,r)}\Big/\binom{M}{\min(j,r)}$ **and the minimum is achieved at** $b=\min(j,r)$**.**

One main critique on the use of the hyper-geometric distribution is the relatively high computational complexity compared to the Gaussian approximation. After all, to get a p-value, one needs to sum all the probabilities from the observed number to the end of the distribution. This may take a long time when the sample size is large. Based on the two properties that (1) the *middle* part of the hyper-geometric distribution, for large sample size, can be approximated very well by a Gaussian distribution and (2) the probability at the *tail* part drops very fast, we propose the following computational scheme, taking the right tail p-value in equation (1) as an example:

$$\begin{aligned} P_R &= \sum_{t=d_s}^{\min(j,r)} p(t;M,j,r) \\ &\approx \begin{cases} G\left(\dfrac{\mu-d_s}{\sigma}\right), & \textbf{if } \mu-d_s \geq 2\sigma \textbf{ and } r \geq 50 \\ \sum_{t=d_s}^{\min(j,r,t_s)} p(t;M,j,r), & \textbf{else} \end{cases} \end{aligned} \tag{3}$$

where $G(\bullet)$ is the standard Gaussian cumulative distribution function, $\mu=\dfrac{j\times r}{M}$, $\sigma^2=\dfrac{jr(M-j)(M-r)}{M^2(M-1)}$, and t_s is the first t such that $p(t;M,j,r)\leq 0.05p(d_s;M,j,r)$.

The above computation scheme maintains the advantages of the hyper-geometric distribution while getting accelerated speed comparable to a Gaussian approximation as indicated in ***Theorem 1***. For the computation associated with large samples, which imply a large number of sum operations, the majority (about 95.4%, which is the probability of $\mathbf{2}\sigma$ in a Gaussian distribution) is replaced by the calculation based on the Gaussian approximation. The rest, including the case with small sample size or the probability on the tail, is computed with an early stopping condition, resulting in a further reduction in the computation time. With t_s chosen as the first t such that $p(t;M,j,r) \le 0.05p(d_s;M,j,r)$, the approximated p-value promises to have at most an error of 5% (***Theorem 1***). It is worth pointing out that the accuracy refers to any point in the distribution, not only for the part with small p-values but also for the part with large p-values.

Theorem 1: **Assume** $\mathrm{H}(b;M,j,r)$ **is the p-value associated with the observed value** b **under the hyper-geometric distribution with parameters** (M,j,r)**. Denote the early-stop estimation as** $\hat{\mathrm{H}}(b;M,j,r) = \sum_{t=b}^{\min(j,r,t_s)} p(t;M,j,r)$ **for** $b \ge \mu + \mathbf{2}\sigma$**, where** $\mu = \frac{j \times r}{M}$, $\sigma^2 = \frac{jr(M-j)(M-r)}{M^2(M-1)}$ **and** t_s **is the first** t **such that** $p(t;M,j,r) \le 0.05p(b;M,j,r)$ **. Then,** $0.95\mathrm{H}(b;M,j,r) \le \hat{\mathrm{H}}(b;M,j,r) \le \mathrm{H}(b;M,j,r)$ **and** $t_s - b \le 1.16\sigma$**.**

II.1.1.F Strategy to re-utilize the existing discoveries

The formulization in equation (1) readily suggests an efficient strategy to take advantage of the previous computations if we are interested (as we often are) in systematically discovering all possible significant SNP subsets ranging from low-order interactions to high-order interactions,

instead of evaluating just one specified SNP subset. After all, the optimization involved in the constraint of equation (1) requires the assessment of all the proper subsets of the SNP subset Φ under investigation. The re-utilization of previous computations will save a lot of time. The seemingly tricky definition of $P_{g\in\phi}=0.05$ for the empty genotype proves to be important to enable us to make use of existing results. Here, 0.05 is chosen because usually we claim an event with a p-value of 0.05 as significant. If a conditional genotype g_m is not empty, we have,

$$\binom{K}{m}\times T\times P_{g_m}\le P_{g\in\phi}=0.05, \tag{4}$$

which means that the corresponding $\mathbf{s}_m$ is significantly associated with the disease after Bonforroni correction. The non-empty conditional genotype should be a significant genotype. Denote Ψ as the set of the significant SNPs discovered in the previous (up to *k* order) search. $\overline{\Psi}\cap\Phi$ stands for the remaining set of SNPs. Here, as long as a SNP is included in a significant SNP subset, we consider that it belongs to Ψ. Thus, to find the significant SNP subsets of (*k*+1)-order, a strategy to utilize the existing results is to divide all the (*k*+1)-order subsets into two groups and handle them differently. The SNP subset in the first group contains member(s) from Ψ, while the SNP subset in the second group does not consist of any member from Ψ. The SNP subset in the first group can be constructed in two steps: (1) choosing *c* SNPs in Ψ with a candidate conditional genotype derived from the previous significant discoveries, where $c=1,...,k$, and (2) choosing *k*+1- *c* SNPs from $\overline{\Psi}\cap\Phi$. The SNP subset in the second group can be obtained in a much easier way, simply choosing $k+1$ SNPs from $\overline{\Psi}\cap\Phi$. Thus, the systematical evaluation of (*k*+1)-order SNP subsets is simplified as solving a constrained optimization problem in the first group with the constraint directly given by previous discoveries and solving an un-constrained optimization problem in the second group. Actually, we can view

the problem for the first group in a new way. If we re-sample the population based on the conditional genotype, that is, the new generated dataset will have all samples possessing the conditional genotype, the constrained optimization problem will be transformed to an unconstrained optimization problem of evaluating k+1- c SNPs from $\overline{\Psi} \cap \Phi$ on the new dataset, supposing the conditional genotype is specified by c SNPs in Ψ. Hence, combining with the problem for the second group, we have a unified unconstrained optimization problem with all the SNPs under examination drawn from $\overline{\Psi} \cap \Phi$, however, possibly based on different datasets. With this new view, the fact that the SNP subset under interrogation is picked up from $\overline{\Psi} \cap \Phi$ has the important and useful implication that none of its proper subsets is significantly associated with the disease and we can utilize this property to further get rid of unnecessary searches, as will be deliberated in our section on heuristic search.

II.1.1.G Approximation to the p-value of P_Φ

P_Φ can be directly used to rank SNP subsets. However, ideally in practice, we also want to know what the p-value associated with such P_Φ is. However, it is very hard to obtain the exact p-value due to the complex dependence structure among the P_g's. An order statistic based on a dependent multivariate Gaussian approximation may offer estimation of the real p-value, but the computation of cumulative distribution function for dependent Gaussian variables is computationally expensive and the estimation is un-reliable for small p-values. More importantly, the tails of the multivariate hyper-geometric distribution are not well approximated by Gaussian multivariates, while the tails are *exactly* what we are interested in.

Denote Φ as a SNP subset with k SNPs, g_m as the conditional genotype specifying the population with M_m subjects, of which, j_m subjects are cases, $M_m - j_m$ subjects are controls. Denote $\mathrm{H}(b; M, j, r)$ as the p-value of the observed b cases based on the hyper-geometric distribution with the parameters M, j and r. From ***Proposition 1***, we know that the minimum p-value of $\mathrm{H}(b; M, j, r)$ over all possible b's is $\mathrm{H}(\min(j, r); M, j, r)$. Define the function $\Gamma(r)$ with independent variable r as

$$\Gamma(r) = \mathrm{H}(\min(j, r); M, j, r). \tag{5}$$

For $P_\Phi \le \Gamma(j)$, define an integer as $d = \Gamma^{-1}(P_\Phi)$ if $\Gamma(d-1) > P_\Phi$, $\Gamma(d) \ge P_\Phi$ and $d \le j$. There exists a unique solution for d, because $\Gamma(r)$ is strictly decreasing when $d \le j$ and $\Gamma(r)$ achieves its minimum at $r = j$ as shown in ***Theorem 2***. Further define $\mathrm{I}(\bullet)$ as an indicator function such that $\mathrm{I}(x \ge y)$ equals 1 if $x \ge y$ is true and 0 otherwise. We propose an efficient yet accurate approximation to the real p-value as follows, with emphasis on more reliable estimation for small p-values:

$$\mathrm{P}(P_\Phi \,\|\, g_m) = 1 - (1 - P_\Phi)^T \tag{6}$$

$$T = \sum_{i=1}^{2^{k-m}} \left(\sum_{j=1}^{2^{k-m}} \mathbf{I}(r_{ij} \ge \Gamma^{-1}(P_\Phi)) \right) \tag{7}$$

where r_{ij} is the number of samples carrying the j^{th} genotype under the i^{th} dichotomy model. Actually, T can be seen as the number of independent multiple tests equivalent to the $2^{2(k-m)}$ dependent tests with boundary effects adjusted.

The approximate p-value specified by (6) and (7) can be derived as follows. As we know from the theory of order statistics, the probability of the smallest value among T independent random variables $\{v_1, v_2, ..., v_T\}$ being less than or equal to P_Φ is,

$$\begin{aligned} &\Pr\left(\min\{v_1, v_2, ..., v_T\} \le P_\Phi\right) \\ &= 1 - \Pr\left(\min\{v_1, v_2, ..., v_T\} > P_\Phi\right) \\ &= 1 - \Pr\left(v_1 > P_\Phi, v_2 > P_\Phi, ..., v_T > P_\Phi\right) \\ &= 1 - \Pr\left(v_1 > P_\Phi\right) \times \Pr\left(v_2 > P_\Phi\right) \times ... \times \Pr\left(v_T > P_\Phi\right) \\ &\approx 1 - \left(1 - P_\Phi\right)^T \end{aligned} \tag{8}$$

The last approximation comes from the fact that $v_i, i = 1, ..., T$ is a p-value and approximately uniformly distributed. From equation (8) we can also clearly see that if the minimum possible value of v_i is larger than P_Φ, then $\Pr\left(v_i > P_\Phi\right) = 1 = \left(1 - P_\Phi\right)^0$ and v_i should be counted as 0 to get the final T. The function $\mathrm{I}\left(r_{ij} \ge \Gamma^{-1}\left(P_\Phi\right)\right)$ in equation (7) makes sure that T only includes tests having the capability to generate a p-value smaller than P_Φ.

The computation of equation (6) requires calculating $\Gamma^{-1}\left(P_\Phi\right)$ for each SNP subset Φ. There is no analytic solution and exhaustive searching over all the possible numbers to find the solution is not efficient because the sample size in GWAS studies is usually quite large. Fortunately, due to the nice properties of $\Gamma(r)$, we have a fast approach, with the computational complexity of logarithm order with respect to the total number of subjects as shown in ***Corollary 2.1***.

Theorem 2: **Assume** $\mathrm{H}(b;M,j,r)$ **is the p-value associated with the observed value** b **under the hyper-geometric distribution of parameters** (M,j,r) **. Define the function** $\Gamma(r)=\mathrm{H}(\min(j,r);M,j,r)$**. Then,** $\Gamma(r)$ **is strictly decreasing for** $r\le j$**, strictly increasing for** $r>j$**, and achieves its minimum at** $r=j$**.**

Corollary 2.1: **There exists a solver of** $\Gamma^{-1}(P_{\Phi})$ **with logarithm computational complexity.**

II.1.1.H Conservativeness of the approximation to the p-value of P_{Φ}

The large number of multiple tests in GWAS studies requires a very stringent experimental-wise significance threshold. In the following analysis we will constrain ourselves to the situation of small p-values. When P_{Φ} is small, we know for the approximated p-value P_{ap} that,

$$P_{ap}=1-(1-P_{\Phi})^{T}\approx T\times P_{\Phi} \qquad (9)$$

Actually, $T\times P_{\Phi}$ is the Bonferroni adjustment involving T multiple tests. It is well known that the Bonferroni adjustment is conservative, that is, the actual p-value is smaller than the one given by the Bonferroni adjustment. The conservativeness becomes more severe when the multiple tests are correlated with each other. We are interested in the degree of the conservativeness for small p-values. The rationale behind this is, if the conservativeness is limited to be of a small degree, we need not bother to appeal to computationally-expensive and sometimes infeasible methods (Genz, 1992) to calculate the accurate p-values.

Let P_{ac} denote the accurate p-value. $R=P_{ac}/(P_{ap})$ is defined to indicate the degree of conservativeness of our proposed approximation. Certainly, R is smaller than or equal to 1 from

the theory of Bonferroni correction. The smaller R is, the more conservative of the approximation. When $R=1$, the approximation equals the true value. Our analysis as in ***Theorem 3*** shows that the degree of the conservativeness is *not* uniform across different true p-values. In fact, R is a decreasing function with respect to the p-value. If we assume that the correlation coefficient is strictly less than 1, we will get that R has the limit of 1 when the p-value approaches to zero. These two properties are good news for our interaction analysis in the context of GWAS studies because GWAS involves a lot of multiple tests and the p-values are required to be small to claim a significant finding. From ***Theorem 3***, we can expect that R is very close to 1 when the p-value is small. These two properties also imply that, without considering the actual level of p-values, it is not appropriate to pursue a uniform correction, such as using PCA to estimate the number of equivalent independent tests and applying this number to adjust the multiple tests.

A lower bound on the conservativeness is also provided by ***Theorem 3***, which allows us to estimate the conservativeness without complicated and extensive computation. The lower bound has the potential application of serving as a ``prejudgment'', to determine whether we need to seek more accurate and yet more expensive approximation. We need to notice that the three conclusions in ***Theorem 3*** are very general and make no particular requirements on the correlation structure of the variables involving in the multiple tests. The first and the third conclusion require no assumption on the correlation structure at all. While we do require that the correlation coefficient be smaller than 1, this requirement is pretty legitimate and an easy preprocessing such as removing identical variables can be performed to satisfy the condition.

Theorem 3 gives the general rules and trends concerning the conservativeness. The actual conservativeness seen in applications will depend on the correlation structure. We have simulated 12 scenarios to get a feeling for the degree of the conservativeness in different cases. The combination of three orders of interaction (two-way to four-way interactions) and four settings of allele frequencies was explored and the empirical values of *R* were recorded. For each scenario one hundred million runs were performed to estimate the true p-value, and thus values for *R*. When the p-value is from 0.05 to 10^{-5}, the estimated *R* has a standard deviation less than 3.16%, where the largest standard deviation is reached at the p-value equal to 10^{-5}. Figure 4 confirms the statements in ***Theorem 3***. *R* is a decreasing function with respect to the p-value and approaches to the limit of 1 when the p-value approaches 0. The lower bound provides a pretty good estimation of *R* and the bound is tight for p-values small enough. Noticing that some empirical *R*s at p-value equal to 10^{-5} are below the lower bound, this is due to the randomness of the empirical *R* because it is estimated from finite simulations. From Figure 4, we can also observe the trends both along the columns and along the rows. Along the columns, the smaller the MAF (minor allele frequency) is, the smaller *R* is, which is expected because the degree of correlation is increased when the MAF is decreased. Along the rows, the larger the order of interaction, the smaller *R* is, which can be explained by the increased number of multiple tests involved.

We are more interested in the degree of conservativeness when the p-value is at the threshold of experimental-wise significance. Note that the threshold of experimental-wise significance will be different for different orders of interaction. A small p-value such as 1e-12 makes the simulation a prohibited task. Instead, in this case, the lower bound of *R* for different orders of interaction is

calculated. Fortunately, the lower bound is pretty tight as illustrated in Figure 4. The results are presented in Table 1, assuming that there are totally 1000 SNPs. The threshold of experimental-wise significance is determined as 0.05 divided by the number of choices for *n* SNPs out of 1000 SNPs, where *n* is the order of interaction. Please be advised that the actual practical conservativeness will be less than the values given in the table since they are the lower bounds. In addition, our approximation will work better than what is shown in the table because in practice usually one examines more than 1000 SNPs, which will result in a more stringent threshold of experimental-wise significance, and, thus, again applying ***Theorem 3*** we know that *R* will be closer to 1 than the corresponding value presented in the table.

From Table 1, we can see that *R* is uniformly good for MAF =0.3 and MAF = 0.5. The observed rule in Figure 4 also applies: smaller MAF implies a smaller R. However, higher order does not imply a smaller R, because the significance threshold is more stringent with higher order. The lowest value of the lower bound is 0.9254, which happens at the three-way interaction and MAF = 0.1. Thus, it appears that we can safely use the simple approximation and avoid expensive computation, since there is only little margin to improve further.

Table 1: The lower bound of *R* when the p-value is at the threshold of experimental-wise significance. Here, we assume there are 1000 SNPs examined and the threshold of experimental-wise significance is determined as 0.05 divided by the number of choices for *n* SNPs out of 1000 SNPs, where *n* is the order of interaction.

Lower Bound for *R*	MAF = 0.1	MAF = 0.3	MAF = 0.5	Mixed MAF
Two-way interaction significance (p-value = 1e-7)	0.9617	0.9957	0.9984	0.9720
Three-way interaction significance (p-value = 3e-10)	0.9254	0.9964	0.9982	0.9430
Four-way interaction significance (p-value = 1.2e-12)	0.9317	0.9988	0.9996	0.9768

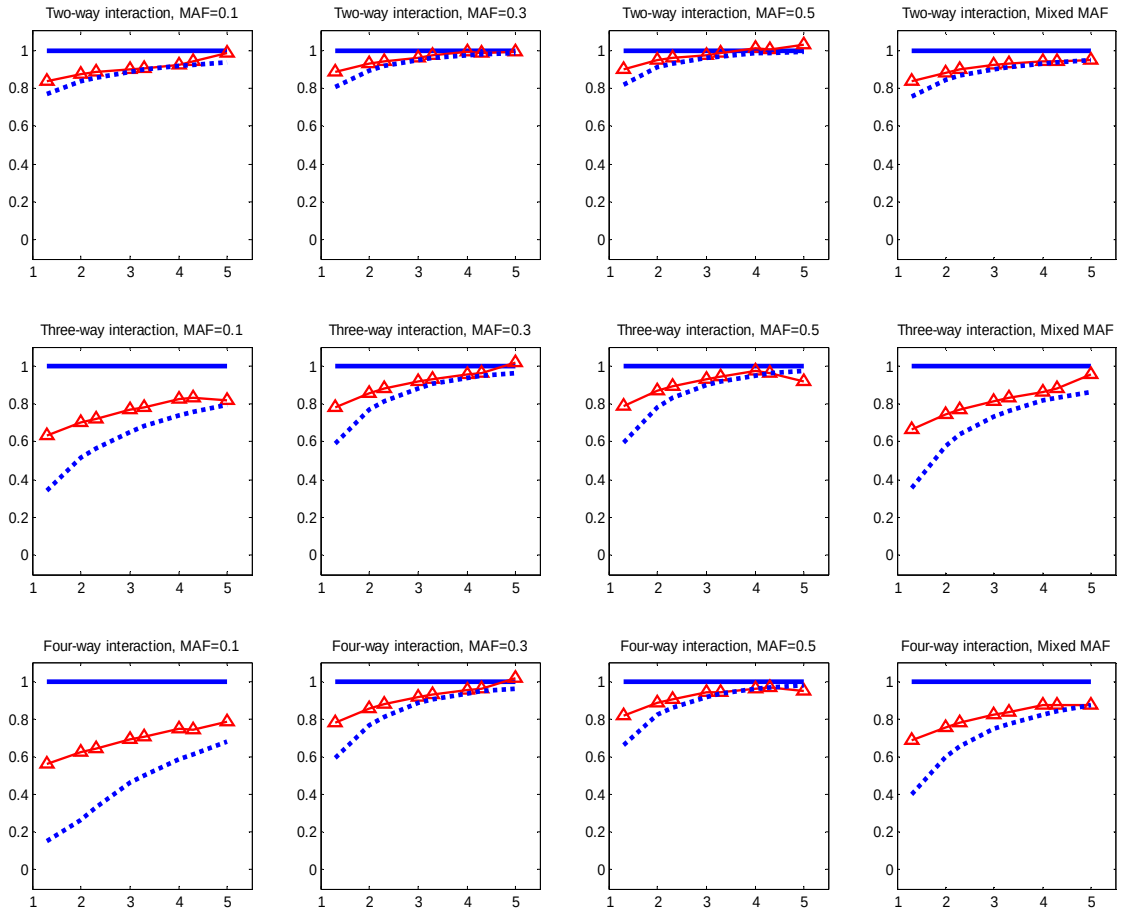

Figure 4: The illustration of the degree of conservativeness for our proposed approximation to the p-values. The X axis indicates the level of significance, which is '-log$_{10}$(p-value)'. The Y axis shows the degree of conservativeness, with the *R* ratio defined as the ratio of the actual p-value over the approximated p-value. The red up-triangle line is the empirical value of *R* obtained on 10^8 simulations. The solid blue line is the limit of R when the p-value approaches zero. The dashed blue line is the lower bound of *R* calculated by our proposed formula in Theorem 3. A total of 12 scenarios were investigated in the simulation study, with the rows corresponding to the three different orders of interaction and the columns corresponding to different settings of minor allele frequency (MAF). MAF = 0.1 means that all the SNPs participating the *n*-way interaction follow the same MAF of 0.1. Similar meanings for MAF = 0.3 and MAF =0.5. Mixed MAF is set to be {0.1, 0.5} for the two-way interaction, {0.1, 0.3, 0.5} for the three-way interaction, and {0.1, 0.2, 0.3, 0.5} for the four-way interaction.

Theorem 3: **Assume the vector** X **of *n* random variables with zero means and unit variances follows a multivariate normal distribution with the covariance matrix as *C*, which has elements** $\{\rho_{ij}\}$**. Denote** $Y = \max\{X_1, X_2, ..., X_n\}$ **and** $\Phi(\bullet)$ **the CDF of the standard normal distribution. Write** $R = \Pr(Y \geq y) \Big/ \left(n(1-\Phi(y))\right)$**. Then, we have: (1), *R* is an increasing function w.r.t. *y*; (2),** $\lim_{y\to\infty} R = 1$ **if** $\rho_{ij} < 1$**; and (3)** $R \geq \frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - \sum_{j=1, j\neq i}^{n} \Phi\left(-y \times \sqrt{(1-\rho_{ij})/(1+\rho_{ij})}\right)\right)$**.**

II.1.1.I Handling missing data

A small fraction of SNP values may be missing in real applications and handling of missing data is thus necessary. One simple choice is to impute missing SNP values randomly according to allele frequencies. However, this approach may inadvertently obfuscate patterns of SNP interaction in some data samples. Actually, any inferring of the missing data values will be biased; thus, an effective strategy is simply to avoid the use of those missing values. Although the missing values are only a small fraction, almost every sample has some SNPs missed. Thus, discarding the samples containing missing values is clearly impractical. Fortunately, our method is based on local counts. Although almost every sample has some SNPs missing, the number of samples with a SNP subset Φ having missing values is expected to be small. Different SNP subsets could have different samples discarded. For instance, one sample with 1000 SNPs may have the 3rd SNP missing. When we evaluate a 2-locus subset not including the 3rd SNP, this sample can be used to count the numbers in equation (1). When the 2-locus subset includes the 3rd SNP, this sample should be discarded.

II.1.1.J Brief summary on incorporating interaction effects for marker discovery

The flow chart for incorporating interaction effects for marker discovery is presented in Figure 1.b. Two key components are the design of the summary statistic score and the significance assessment of the observed score. The score is obtained from a constrained combinatorial optimization problem, in which P_s, the p-value computed from the hypergeometric distribution, plays an important role. The accurate assessment of significance is non-trivial and several characteristics unique to GWAS have to be taken into account, such as the correlation structure among genotypes, the cross-order comparison, the discreteness and the finitesimality of hypergeometric distribution based p-values. It is worth noting that the score reduces to Fisher's exact test when it is applied to individual SNP analysis.

In the process of devising the approach of incorporating interaction effects for marker discovery, we made the follow mathematical contributions. (1) We have proved the non-infinitesimal characteristic of the hyper-geometric distribution and proposed an effective solution to correct the effect. (2) We have developed a fast approach to evaluate the hyper-geometric distribution and show that the proposed approach is within allowable inaccuracy. (3) We have investigated the characteristics of the correlation structure's influence on the significance assessment and identified the conditions that offer quite an easy and simple solution. Fortunately, our problem in satisfies these conditions.

II.1.2 Identifying Interaction Effects among Significant SNPs

This subsection is devoted to the second task, involving the interaction effects in GWAS, that is, the identification of interaction effects among marginally significant SNPs. First, we define "biological interaction'' in subsubsection II.1.2.A. This definition inspires us to propose the test statistic as discussed in subsubsection II.1.2.B. The assessment of significance is a direct application of Wilks' theorem for the likelihood ratio test. However, the maximum likelihood estimation associated with the test is not straightforward and it is subject to a finite but large set of linear inequalities. In subsubsection II.1.2.C we design a two stage approach to this constrained maximum likelihood estimation problem by combining the Newton method and the barrier method. In subsubsection II.1.2.D we further investigate several critical mathematical properties that have great practical significance. These include "compatibility'' of the model (to be defined in the sequel) with the existence of unmeasured disease factors, with linkage disequilibrium structure (and, more generally, with surrogate measured factors), and under disease heterogeneity (the case of multiple disease subtypes). As the logistic regression with interaction terms is the most popular, widely used approach for this problem, we have mathematically compared its power to that of our model, establishing the superiority of our method in subsubsection II.1.2.E.

As we will see the discussions followed, we figure out three key properties a detection criterion should ideally possess, in trying to identify interactions with significant marginal effects: the compatibility with 1) disease heterogeneity; 2) hidden risk factors and 3) tag markers. We will mathematically prove that our proposed criterion satisfies these three criteria, while, surprisingly, conventional logistic regression with interaction terms does not satisfy any of these three

requirements. We will also rigorously prove that our approach is more powerful than conventional logistic regression with interaction terms.

II.1.2.A Definition of biological interaction

Interaction effects have many implications both in GWAS studies and in scientific case-control studies in general. Here, our main focus will be on GWAS. Besides improving the power to detect novel markers associated with disease by incorporating interaction effects, the mere justification of existence of interaction among multiple SNPs with significant marginal effects is interesting and important to the disease mechanism inference and to the design of efficient preventive and/or disease treatment plans. At the same time, the identification of interaction among significant SNPs is an integral part of our SCA framework, as it will help extract conditional genotypes from existing significant SNPs to re-utilize the existing computation.

Logistic regression with multiplicative interaction terms (Agresti, 2002) is probably the most popular tool available. Its dominance is mainly due to its mathematical simplicity, familiarity, and its great availability in standard software packages to practitioners in the field. However, there is no justification for the specific interaction model it uses and there is no direct relation to biological interactions, as we know statistical interaction and biological interaction are usually two different things. Here, we propose a new measure for interaction effects with direct link to a plausible biological model for disease acquisition.

The interaction in biology is complex and complicated. For example, multiple factors may work in parallel, serial, or mixed. In the parallel sense, it means the robustness of biology or the redundancy, that is, if one factor does not work, the other factors can still work. The gene pairs of synthetic lethality work in a parallel fashion. On the other hand, in the serial sense, it is like a chain, such that if one factor fails the whole chain breaks. Actually, we believe biology usually works in a mixed way. Although the interaction could have various forms, there is only one way to ``name'' the non-interaction effect to the disease. When we say multiple disease factors jointly influence the disease risk without interaction, we mean these factors influence the disease risk independently of each other. However, it is important to distinguish the subtle difference between the "independence" used here and the concept of independence in probability theory. The two concepts are certainly closely related, since independence in probability describes the relationship among random events. However, we need to pinpoint the event of interest. Does the event represent "being diseased" or "being normal"? The distinction makes a big difference. Indeed, the independence here concerns "being normal", not "being diseased", although we are talking about "disease"-risk factors.

Suppose there are M disease factors, F_1, F_2, …, F_{M},. Assume the disease factor is binary, with 0 representing the normal status and 1 the risky status. Corresponding to each disease factor F_{m}, there is a probability φ_m measuring the likelihood of being diseased if a (heretofore healthy) subject has the disease factor changed from normal status to risky status. The probability φ_m can be understood from the standpoint of the following virtual experiment: Choose one healthy person with $F_m = 0$, make $F_m = 1$ and let the person re-start life with all other factors unchanged. Then, how likely is it for this person to get the disease? Certainly, it is impossible to implement

the experiment. But it is not a forged concept. Indeed, we can look at the population and estimate φ_m by comparing the likelihood of disease with $F_m = 1$ to the one with $F_m = 0$. Denote η_{m0} as the probability of disease cases for $F_m = 0$. Similarly, Denote η_{m1} as the probability of disease cases for $F_m = 1$. Thus, φ_m can be estimated as,

$$\varphi_m = 1 - \frac{1-\eta_{m1}}{1-\eta_{m0}} \tag{10}$$

Let φ_0 denote the probability of being diseased when one is free of all the *M* disease factors, that is when $F_m = 0$ for $m = 1,...,M$. φ_0 can be considered as the background probability of disease, due to various reasons unaccounted for by the *M* disease factors. For complex diseases, there are multiple disease causes or factors. If any one of the factors unfortunately takes effect and, further, if the presence of this factor causes the human system to fail to work, the person will be diseased. On the other hand, for a person to be healthy, it requires that none of the disease factors that is present actually causes the disease to occur. Thus, under the assumption of no interaction, for a subject with disease factors $F_m = f_m$, the probability of disease is,

$$\Pr\left(dis \mid F_1 = f_1, F_2 = f_2, ..., F_M = f_M\right) = 1 - \left(1-\varphi_0\right)\prod_{m=1}^{M}\left(1-\varphi_m\right)^{f_m} \tag{11}$$

where f_m can take values of 0 or 1, for $m = 1,...,M$.

The multiplication of probabilities in the right side of equation (11) comes from the fact that all the disease factors need to output healthy status for a person to be normal and from the fact that the factors work independently. Equation (11) can be read and simulated in a stochastic process as follows. Given a person with disease factors $F_m = f_m$ for $m = 1,...,M$, at the very beginning

initialize the person's status with the probability $(1-\varphi_0)$ being healthy, if the status is diseased, then the process is stopped and the final status of the person is diseased. Otherwise, continue to set the person's status with the probability $(1-\varphi_1)^{f_1}$ being healthy. Generally, if the current step involves the factor F_m and the status is diseased, stop the process and the final status is diseased. If the current status is healthy, continue to set the next status with the probability $(1-\varphi_{m+1})^{f_{m+1}}$ being healthy. Therefore, for a person to be healthy (as the final status, after all the *M* factors are examined), at each step its status should be set as healthy, *i.e.* any disease factors that are in fact present do not (by chance) cause the disease. According to the definition of the probability φ_m, the probability of the final status being healthy is obtained by multiplying each step's healthy probability together, that is, $(1-\varphi_0)\prod_{m=1}^{M}(1-\varphi_m)^{f_m}$. Naturally, the disease probability follows as in equation (11). Note that φ_m is defined solely based on the disease factor F_m, without dependence on other factors. The independence in the assumption of non-interaction is why we simply multiply together all the individual factor probabilities.

II.1.2.B Test statistic for interaction effect among disease-risk SNPs

Interaction is defined as a deviation from the model for *non-interaction* described by equation (11). Suppose we have *M* SNPs, $X_1, X_2, ..., X_M$, which are associated with the disease. It is easy to determine the disease-risk allele. Further, we impose a dominant/recessive model to the SNP, such that X is binary. Trivially, we can transform the SNP data so that $X=0$ decodes to normal and $X=1$ decodes to risky. Hence, under the null hypothesis of non-interaction, corresponding to equation (11), we have

$$\log\left(1-\Pr\left(dis\right)\right)=\alpha+\sum_{m=1}^{M}x_m\beta_m \tag{12}$$

We are interested in the genotype that all X are 1. We call it the 'full genotype', denoted by X_{full}. It is expected that if there are interactions among *X*s, it is most likely the genotype X_{full} that will account for a disease probability that differs from the one predicted by the null hypothesis. So we insert in one additional variable specifying the full genotype. Therefore, under the alternative hypothesis, we have,

$$\log\left(1-\Pr\left(dis\right)\right)=\alpha+\sum_{m=1}^{M}x_m\beta_m+x_{full}\beta_{full} \tag{13}$$

$$\text{where } x_{full}=\begin{cases}1 & x_1=x_2=\ldots=x_m=1\\ 0 & otherwise\end{cases}$$

Denote the maximum log-likelihood under the null hypothesis as L_0. Denote the maximum log-likelihood under the alternative hypothesis as L_1. We propose the test statistic to measure the deviation as follows,

$$G=2\left(L_1-L_0\right) \tag{14}$$

The null model can be seen as a special case of the alternative model, achieved by forcing $\beta_{full}=0$. L_0 will never be larger than L_1, because, if L_0 is larger than L_1, the special case $\beta_{full}=0$ will result in a larger log-likelihood based on the alternative model, which contradicts with the fact that L_1 is the maximum log-likelihood under the alternative model. According to **Wilks' theorem** (Wilks, 1938), *G* asymptotically follows a chi-squared distribution with 1 degree of freedom.

Unlike the logistic regression model, which assumes the linear form for the logarithm of the odds ratio, we take the linear form for the logarithm of the probability of being *healthy*. This is an important difference, both conceptually and computationally. Because the logistic regression model considers the disease status and the healthy status symmetrically, it will not affect the final results if we swap the labels of case and control. Our model considers the disease status and the healthy status having different underlining mechanisms. It will get wrong results if the labels are swapped because the logarithm of the probability of disease does not have a linear form. In addition, the logistic regression model optimizes the log-likelihood over the full parameter space. By contrast, in our model, notice that $\Pr(dis) \geq 0$, so we have $\alpha + \sum_{m=1}^{M} x_m \beta_m \leq 0$ and $\alpha + \sum_{m=1}^{M} x_m \beta_m + x_{full} \beta_{full} \leq 0$. These constraints on the parameters (one for every data point in the given case-control population) require us to design a special maximum likelihood estimation algorithm, as described in the next subsection.

When $\Pr(dis)$ approximates to 1 and $\log(\Pr(dis)) \to 0$, we have,

$$\log\left(\Pr(dis)/(1-\Pr(dis))\right) \approx -\log(1-\Pr(dis)) \qquad (15)$$

In this case, our model and logistic regression converge to the same model. However, as the control sample is much easier to collect in GWAS studies, the fraction of the cases in studies is usually close to 0, rather than close to 1. Thus, the models are different in practice, and it is possible that we may be able to find true interactions in real applications through our model, while, for the same problem, the logistic regression model fails. In fact, we will demonstrate that

our model is both more powerful in general and that indeed it can detect significant interactions on real GWAS data sets for which logistic regression fails to detect.

We propose a progressive (sequential) dissection on the interaction effects to avoid the contamination of a small interacting subset on the larger subset. Accordingly, when an *M*-SNP subset is determined to be significant, we construct a pseudo-SNP to replace this subset, with 1 representing the full genotype as defined in equation (13) and 0 representing all the other genotypes.

II.1.2.C Constrained maximum likelihood estimation

The standard approach to maximum likelihood estimation for generalized linear models (McCullagh and Nelder, 1989) cannot be applied here because we have constraints on the parameter space. Thus, we design a special optimization algorithm to cope with the constraints. Since the alternative model in equation (13) can be thought of as a null model with $M+1$ SNPs as in equation (12), for simplicity of illustration, we only show how to estimate parameters in the null model.

Suppose there are *N* subjects with l_i denoting the label, where $l_i = 0$ indicates that the *i*th subject is a control and $l_i = 1$ indicates that the i^{th} subject is a case. Let the *M*-dimensional column vector X_i denote the i^{th} subject's genotype with each element corresponding to one SNP. For concision of expression, we use the symbols $Y_i = [1, X_i^T]^T$ and $\theta = [\alpha, \beta_i^T]^T$. Then according to equation (12), the probability for the i^{th} subject to be a case is $1 - e^{\theta^T Y_i}$. Correspondingly, the probability

for the i^{th} subject to be a control is $e^{\theta^T Y_i}$. Thus, given the parameter vector θ, the log-likelihood for all the *N* subjects is

$$L(\theta) = \sum_{i=1}^{N} \left((1 - l_i)\theta^T Y_i + l_i \log\left(1 - e^{\theta^T Y_i}\right) \right) \quad . \quad (16)$$

The maximum likelihood estimation of θ can be formalized as the following optimization problem:

$$\max_{\theta} L(\theta) \quad (17)$$

$$subject\ to: \theta^T Y_i \le 0,\ for\ i = 1,...N$$

The constraint in equation (17) comes from the fact that the probability should be in the range $[0,1]$, that is, $0 \le e^{\theta^T Y_i} \le 1$. Then, we have $\theta^T Y_i \le 0$. Although there seems to be N constraints in equation (17), some constraints may be duplicated and there are at most 2^M different constraints.

From ***Theorem 4*** we know that $L(\theta)$ is a concave function. In addition, the inequality constraints in equation (17) are all linear. So the identical minimization problem $\min_{\theta}(-L(\theta))$ is a convex problem (Boyd and Vandenberghe, 2004), such that reliable and efficient optimization approaches are available. With the observation that in most cases the solution in equation (17) is equivalent to the corresponding unconstrained problem, we propose a mixed approach that takes advantage of the unconstrained optimization problem before proceeding to the constrained problem. The algorithm has two parts. In the first part, we use the **Newton method** (Boyd and Vandenberghe, 2004) to solve the unconstrained problem, with a check on the validation of the constraints at each iteration. If any constraint is violated before convergence, we stop this part

and go to the second part; otherwise, the second part will not be called and the solution of this part is the solution to the original problem. In the second part, we use the **barrier method** (Boyd and Vandenberghe, 2004) to solve the constrained problem. The basic idea is to transform the constrained problem into a sequence of unconstrained problems indexed by the parameter *t*. When *t* goes to infinity the corresponding unconstrained problem approximates, to infinitesimal precision, the original constrained problem. Given a certain *t*, we apply the Newton method again to solve the unconstrained problem.

First part: Newton method to solve the unconstrained problem

Given the current estimate $\theta^{(k)}$ of parameter θ, the next estimate $\theta^{(k+1)}$ is updated as in the following:

$$\theta^{(k+1)} = \theta^{(k)} + \mathrm{A}\left(\theta^{(k)}\right) \times \mathrm{B}\left(\theta^{(k)}\right) \tag{18}$$

where,

$$\mathrm{A}\left(\theta^{(k)}\right) = -\frac{\partial^2}{\partial \theta^2} L\left(\theta^{(k)}\right) = \sum_{i=1}^{N} \left(l_i \frac{e^{\theta^{(k)T} Y_i}}{\left(1 - e^{\theta^{(k)T} Y_i}\right)^2} Y_i Y_i^T \right) \tag{19}$$

$$\mathrm{B}\left(\theta^{(k)}\right) = \frac{\partial}{\partial \theta} L(\theta) = \sum_{i=1}^{N} \left((1 - l_i) Y_i + l_i \frac{e^{\theta^{(k)T} Y_i}}{1 - e^{\theta^{(k)T} Y_i}} Y_i \right). \tag{20}$$

We initialize $\theta^{(0)}$ with the first element $\theta_1^{(0)} = -0.6931$ and all the other elements $\theta_i^{(0)} = 0$ for $i = 2, ..., M+1$. The stop criterion is set as $\left| L\left(\theta^{(k+1)}\right) - L\left(\theta^{(k)}\right) \right| \le 10^{-6}$. The Newton method is a very powerful technique with quadratic convergence rate. It usually converges in less than 10 iterations in our application.

Second part: Barrier method to solve the constrained problem

The idea for the barrier method is simply to transform the constrained problem into an unconstrained problem. Recall that our constrained problem is as follows:

$$\max_{\theta} L(\theta)$$

$$subject\ \ to: \theta^T Y_i \le 0,\ for\ i = 1,...N\ .$$

Denoting that $\mathrm{I}_+(x) = \begin{cases} 0, & \textbf{if } x > 0 \\ -\infty, & \textbf{if } x \le 0 \end{cases}$, then the preceding constrained optimization problem is equivalent to the following problem:

$$\max_{\theta} L(\theta) + \sum_{i=1}^{N} \mathrm{I}_+(-\theta^T Y).$$

We can easily check that the optimum solution must occur at $\theta^T Y_i \le 0,\ for\ i = 1,...N$; otherwise the objective function will be negative infinity and this will clearly not be an optimum solution. We can also see that the optimum value of the modified objective function is also equal to an optimal value of the original objective function, since $\mathrm{I}_+(x) = 0$ for feasible solutions.

However, the difficulty is that the above problem cannot be solved directly due to the non-differentiability. The barrier method proposes a new scheme to replace $\mathrm{I}_+(x) = 0$ by an approximated function $\varphi_t(x)$ that is differentiable, where the accuracy of the approximation can be controlled by the parameter t>0. A good candidate for $\varphi_t(x)$ is formulated as the following logarithmic barrier function:

$$\varphi_t(x)=\begin{cases}\frac{1}{t}\log(x) & \textbf{if } x>0\\ \text{-}\infty & \textbf{if } x\le 0\end{cases}.$$

It is easy to show that

$$\varphi_t(x)\xrightarrow{t\to\infty}\mathrm{I}_+(x)=\begin{cases}0, & \textbf{if } x>0\\ -\infty, & \textbf{if } x\le 0\end{cases}.$$

Given a certain *t*, we solve the following unconstrained problem:

$$\max_\theta\left\{L(\theta)+\sum_{i=1}^{N}\frac{1}{t}\log\left(-\theta^T Y_i\right)\right\} \tag{21}$$

It can be easily verified that this is a concave problem which promises the global optimization. For each *t*, we apply the Newton method and get the update rule that is the same as in equation (18), but with the $\mathrm{A}\left(\theta^{(k)}\right)$ and $\mathrm{B}\left(\theta^{(k)}\right)$ replaced by,

$$\begin{aligned}\mathrm{A}\left(\theta^{(k)}\right)&=-\frac{\partial^2}{\partial\theta^2}\left(L\left(\theta^{(k)}\right)+\sum_{i=1}^{N}\frac{1}{t}\log\left(-\theta^T Y_i\right)\right)\\&=\sum_{i=1}^{N}\left(\left(l_i\frac{e^{\theta^{(k)T}Y_i}}{\left(1-e^{\theta^{(k)T}Y_i}\right)^2}+\frac{1}{\left(\theta^T Y\right)^2}\right)Y_iY_i^T\right)\end{aligned} \tag{22}$$

$$\begin{aligned}\mathrm{B}\left(\theta^{(k)}\right)&=\frac{\partial}{\partial\theta}\left(L\left(\theta^{(k)}\right)+\sum_{i=1}^{N}\frac{1}{t}\log\left(-\theta^T Y_i\right)\right)\\&=\sum_{i=1}^{N}\left(\left(1-l_i\right)Y_i+l_i\frac{e^{\theta^{(k)T}Y_i}}{1-e^{\theta^{(k)T}Y_i}}Y_i+\frac{1}{\theta^T Y_i}Y_i\right)\end{aligned} \tag{23}$$

As t goes to positive infinity, the modified problem in equation (21) approaches the original problem.

The control parameter is updated as,

$$t^{(k+1)} = \lambda t^{(k)},$$

where $\lambda > 1$. λ controls the speed of increase of t. Call the iteration for t as the outer loop and the iteration of the Newton method for a given t as the inner loop. While a large λ is desirable to reduce the outer loop iteration, it also increases the number of the inner loop iterations. There is thus a tradeoff in the choice of λ. Both theoretical analysis and empirical experiments show that values from 3 to 100 seem to be good choices. In our experiments we set $\lambda = 20$.

Suppose that the constraints are violated at the K^{th} iteration in the Newton method on the original unconstrained problem. Then we have the valid increase of log-likelihood as $V = L\left(\theta^{(K-1)}\right) - L\left(\theta^{(K-2)}\right)$. With the guidance that the initialization of $t^{(0)}$ performs very well when $N/t^{(0)}$ is close to $\lambda\left(L^* - L\left(\theta^{(0)}\right)\right)$, where L^* is the true maximum, we choose that $t^{(0)} = N/(\lambda V)$, with the expectation that V reasonably approximates to $\left(L^* - L\left(\theta^{(0)}\right)\right)$.

There is another initialization concerning the inner loop. This is pretty straightforward and actually the source of the strength of the barrier method, that is, the initial θ for the Newton method at $t^{(k+1)}$ is simply the final θ at $t^{(k)}$.

There are two stopping criterion, one for the outer loop and one for the inner loop. For the inner loop, we stop if the increase in log-likelihood is less than 10^{-6}. For the outer loop, we stop if $t^{(k)} > N \times 10^6$. From the theory of the dual problem (Boyd and Vandenberghe, 2004), the stop criteria promise that the error between the modified problem in equation (17) and the original problem in equation (21) is less than 10^{-6}.

Theorem 4: $L(\theta)$ **is a concave function.**

II.1.2.D Mathematical properties of the proposed statistic

Besides the theoretical-sounding significance assessment and the nice convex property of the corresponding maximum likelihood estimation, our model enjoys three critical properties that are desirable for a model to be highly successful in real applications. First, the model should be compatible when some disease-risk factors are not measured or hidden. Second, the model should be compatible when the measured/observed factors are correlated to but are not in fact the causal factors. Third, the model should be compatible in the situation where the disease is heterogeneous (to be defined in the sequel). In contrast to our model, the logistical regression strikingly does not satisfy *any* of the three properties.

Our new model form still holds when some disease-causing factors are not measured. For comparison, the logistic regression model does not hold for this very common situation. This property is rigorously established through ***Lemma 5.1*** and ***Theorem 5***. It should be pointed out this property is very important for practical purposes since we rarely have the luxury to measure all the potential disease-causing factors. Even if we do have the luck to observe all the factors,

when we interrogate the interaction effect among some (but not all) factors, it is equivalent to having other factors missed.

Will the new model still hold when the measured markers are not causal but are in fact tag markers? The quick answer is yes. For comparison, the logistic regression model does not hold for this very typical situation. This property is rigorously established through ***Lemma 6.1***, ***Theorem 6***, ***Corollary 6.1*** and ***Corollary 6.2***. In GWAS studies, the discovered disease-risk SNPs are usually tag-SNPs that are in high linkage disequilibrium with the causal DNA mutations. Here, we should point out the following caveat. Even if the causal DNA mutation (for example, a SNP) is a linear function of the number of alleles, that is, the effect of genotype 2 is two times the effect of genotype 1, this linear relationship is *destroyed* and does not hold generally for the observed tag-SNPs. Fortunately, no matter how complicated the relation of effects among different number of alleles in one mutation site is and how intricate the correlation structure is for the tag-SNPs, our proposed model form *always holds* for dominant/recessive genetic coding, that it remains invariant when there are unmeasured causal SNPs or tag SNPs.

If the disease is heterogeneous and we assume that each of its subtypes follows the new model, the new model still holds for the disease, while the logistic regression model does not hold for this situation. This property is rigorously established through ***Theorem 7***. This property is very basic for the success of a case-control study on a complex disease because the complex disease is usually heterogeneous.

To give concrete illustration on how these mathematical yet practically desirable properties hold for the new model and are violated for the logistic regression model, we provide three sets of examples. For the sake of simplicity, we assume that each SNP has two statuses of 0 and 1 in the following discussion.

Example 1.a on the missing risk factors of the logistic regression model:

In this example, we assume that three SNPs (X1, X2 and X3) contribute to the disease risk under the logistic regression model described by the following formula,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where p is the disease probability associated with each genotype.

Taking the parameter setting that $\beta_0 = -4$, $\beta_1 = 2$, $\beta_2 = 2$ and $\beta_3 = 2$, we have Table 2 showing the disease probability under each genotype.

Table 2: Distribution of disease probability with all the risk factors observed under the logistic regression model

X3 = 0			X3 = 1		
p	X2 = 0	X2 = 1	p	X2 = 0	X2 = 1
X1 = 0	0.0180	0.1192	X1 = 0	0.1192	0.5000
X1 = 1	0.1192	0. 5000	X1 = 1	0.5000	0. 8808

Suppose the third risk factor X3 is missing. We further assume $\Pr(X_3 = 0) = 0.5$ and $\Pr(X_3 = 1) = 0.5$. Then the distribution of disease probability with the third risk factor X3

missing can be obtained by averaging the left and the right sub-tables of Table 2. The new distribution is shown in Table 3.

Table 3: Distribution of disease probability with the third risk factor X3 missing under the logistic regression model

p	X2 = 0	X2 = 1
X1 = 0	0.0686	0.3096
X1 = 1	0.3096	0.6904

If the new disease model holds for the case that X3 is missing, denoting $\beta_0^{'}$, $\beta_1^{'}$ and $\beta_2^{'}$ the parameters for the new logistic model, we can compute these parameters based on the upper-left three genotypes in Table 3 by the following,

$$\beta_0^{'} = \log\left(\frac{\Pr(p \mid X_1 = 0, X_2 = 0)}{1 - \Pr(p \mid X_1 = 0, X_2 = 0)}\right) = \log\left(\frac{0.0686}{1 - 0.0686}\right) = -2.6084$$

$$\beta_1^{'} = \log\left(\frac{\Pr(p \mid X_1 = 0, X_2 = 1)}{1 - \Pr(p \mid X_1 = 0, X_2 = 1)}\right) - \beta_0^{'} = \log\left(\frac{0.3096}{1 - 0.3096}\right) + 2.6084 = 1.8064$$

$$\beta_2^{'} = \log\left(\frac{\Pr(p \mid X_1 = 1, X_2 = 0)}{1 - \Pr(p \mid X_1 = 1, X_2 = 0)}\right) - \beta_0^{'} = \log\left(\frac{0.3096}{1 - 0.3096}\right) + 2.6084 = 1.8064$$

So, we can calculate the odds for the genotype (X1 = 1, X2 = 1) based on the logistic regression model as the following,

$$\frac{\Pr(p \mid X_1 = 1, X_2 = 1)}{1 - \Pr(p \mid X_1 = 1, X_2 = 1)} = e^{\beta_0^{'} + \beta_1^{'} \times 1 + \beta_1^{'} \times 1} = 2.7385$$

However, from Table 3 we know the odds for the genotype (X1 = 1, X2 = 1) is $\frac{0.6904}{1-0.6904} = 2.2300$. So the logistic regression model breaks (its form is not invariant) when some risk factors are missing.

Example 1.b on the missing risk factors of the new model:

In this example, we assume that three SNPs (X1, X2 and X3) contribute to the disease risk under the new model described by the following formula,

$$\log(1-p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where *p* is the disease probability associated with each genotype.

Taking the parameter setting that $\beta_0 = -0.0182$, $\beta_1 = -0.1088$, $\beta_2 = -0.1088$ and $\beta_3 = -0.1088$, we have Table 4 showing the disease probability under each genotype.

Table 4: Distribution of disease probability with all the risk factors observed under the new model

X3 = 0			X3 = 1		
p	X2 = 0	X2 = 1	p	X2 = 0	X2 = 1
X1 = 0	0.0180	0.1192	X1 = 0	0.1192	0.2100
X1 = 1	0.1192	0. 2100	X1 = 1	0.2100	0.2914

Suppose the third risk factor X3 is missing. We further assume $\Pr(X_3 = 0) = 0.5$ and $\Pr(X_3 = 1) = 0.5$. Then the distribution of disease probability with the third risk factor X3 missing can be obtained by averaging the left and the right sub-tables of Table 4. The new distribution is shown in Table 5.

Table 5: Distribution of disease probability with the third risk factor X3 missing under the new model

p	X2 = 0	X2 = 1
X1 = 0	0.0686	0.1646
X1 = 1	0.1646	0.2507

On one hand, if the new disease model holds for the case that X3 is missing, denoting β_0', β_1' and β_2' the parameters for the new logistic model, we can compute these parameters based on up-left three genotypes in Table 5 by the following,

$$\beta_0' = \log\left(1 - \Pr\left(p \mid X_1 = 0, X_2 = 0\right)\right) = \log\left(1 - 0.0686\right) = -0.0711$$

$$\beta_1' = \log\left(1 - \Pr\left(p \mid X_1 = 0, X_2 = 1\right)\right) - \beta_0' = \log\left(1 - 0.1646\right) + 0.0711 = -0.1088$$

$$\beta_2' = \log\left(1 - \Pr\left(p \mid X_1 = 1, X_2 = 0\right)\right) - \beta_0' = \log\left(1 - 0.1646\right) + 0.0711 = -0.1088$$

So, we can calculate the odds for genotype (X1 = 1, X2 = 1) based on the new model as the following,

$$\frac{\Pr\left(p \mid X_1 = 1, X_2 = 1\right)}{1 - \Pr\left(p \mid X_1 = 1, X_2 = 1\right)} = \frac{1 - e^{\beta_0' + \beta_1' \times 1 + \beta_1' \times 1}}{e^{\beta_0' + \beta_1' \times 1 + \beta_1' \times 1}} = 0.3346 .$$

On the other hand, from Table 2 we know the odds for genotype (X1 = 1, X2 = 1) is $\frac{0.2507}{1 - 0.0.2507} = 0.3346$, which empirically confirms the theorem on missing factors for the new model.

Example 2.a on measuring tag SNPs for the logistic regression model:

In this example, we have two causal SNPs (X1 and X2) that follow the logistic regression model,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .$$

Assuming the parameters $\beta_0 = -4$, $\beta_1 = 2$ and $\beta_2 = 2$, we can get Table 6 showing the disease probability under each genotype.

Table 6: Distribution of disease probability with two causal SNPs X1 and X2 under the logistic regression model

p	X2 = 0	X2 = 1
X1 = 0	0.0180	0.1192
X1 = 1	0.1192	0. 5000

Suppose two tag SNPs (X1' and X2') are measured instead. The correlation structure between X1' and X1 is given by the conditional probabilities,

$$\Pr\left(X_1' = 0 \mid X_1 = 0\right) = 0.9,$$

$$\Pr\left(X_1' = 1 \mid X_1 = 0\right) = 0.1,$$

$$\Pr\left(X_1' = 0 \mid X_1 = 1\right) = 0.1,$$

$$\Pr\left(X_1' = 1 \mid X_1 = 1\right) = 0.9.$$

Assuming the probabilities $\Pr\left(X_1 = 0\right) = 0.5$ and $\Pr\left(X_1 = 1\right) = 0.5$, we find that the correlation coefficient between X1' and X1 is 0.8.

Similarly, the correlation structure between X2' and X2 is set by,

$$\Pr\left(X_2' = 0 \mid X_2 = 0\right) = 0.9,$$

$$\Pr\left(X_2^{'}=1 \mid X_2=0\right)=0.1,$$

$$\Pr\left(X_2^{'}=0 \mid X_2=1\right)=0.1,$$

$$\Pr\left(X_2^{'}=1 \mid X_2=1\right)=0.9.$$

Using the assumption that X1 and X2 are independent and X1' and X2' are independent, we can get the disease distribution associated with the tag genotypes in terms of the ones corresponding to the causal genotypes,

$$\begin{aligned}
&\Pr\left(p \mid X_1^{'}=k_1, X_2^{'}=k_2\right) \\
&=\frac{\Pr\left(p, X_1^{'}=k_1, X_2^{'}=k_2\right)}{\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2\right)} \\
&=\frac{\sum_{i=0}^{1}\sum_{j=0}^{1}\Pr\left(p, X_1^{'}=k_1, X_2^{'}=k_2, X_1=i, X_2=j\right)}{\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2\right)} \\
&=\frac{\sum_{i=0}^{1}\sum_{j=0}^{1}\Pr\left(p, X_1^{'}=k_1, X_2^{'}=k_2 \mid X_1=i, X_2=j\right)\Pr\left(X_1=i, X_2=j\right)}{\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2\right)} \\
&=\frac{\sum_{i=0}^{1}\sum_{j=0}^{1}\Pr\left(p, \mid X_1=i, X_2=j\right)\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2 \mid X_1=i, X_2=j\right)\Pr\left(X_1=i, X_2=j\right)}{\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2\right)}
\end{aligned}.$$

The last equation holds due to the fact that the genotypes of the tag SNPs are determined by the causal SNPs, irrespective to the disease status. In this special case we can see that $\Pr\left(X_1=i, X_2=j\right)=0.25$ and $\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2\right)=0.25$. The calculation can be reduced to,

$$\Pr\left(p \mid X_1^{'}=k_1, X_2^{'}=k_2\right)=\sum_{i=0}^{1}\sum_{j=0}^{1}\Pr\left(p, \mid X_1=i, X_2=j\right)\Pr\left(X_1^{'}=k_1, X_2^{'}=k_2 \mid X_1=i, X_2=j\right).$$

Table 7: Distribution of disease probability with two tag SNPs X1' and X2' under the logistic regression model

p	X2' = 0	X2' = 1
X1' = 0	0.0410	0.1444
X1' = 1	0.1444	0.4266

Let us show for one instance how the above formula can be used to derive the disease distribution under the genotype specified by the tag SNPs:

$$\begin{aligned}&\Pr\left(p \mid X_1^{'}=0, X_2^{'}=0\right)\\ &=0.0180*0.9*0.9+0.1192*0.9*0.1+0.1192*0.9*0.1+0.5000*0.1*0.1\\ &=0.0410\end{aligned}$$

Then the disease distribution on each genotype specified by the tag SNPs is summarized in Table 7. Again, denoting $\beta_0^{'}$, $\beta_1^{'}$ and $\beta_2^{'}$ the parameters for the logistic model with the tag SNPs, we can compute these parameters based on Table 7 by the following,

$$\beta_0^{'}=\log\left(\frac{\Pr\left(p \mid X_1^{'}=0, X_2^{'}=0\right)}{1-\Pr\left(p \mid X_1^{'}=0, X_2^{'}=0\right)}\right)=\log\left(\frac{0.0410}{1-0.0410}\right)=-3.1523,$$

$$\beta_1^{'}=\log\left(\frac{\Pr\left(p \mid X_1^{'}=0, X_2^{'}=1\right)}{1-\Pr\left(p \mid X_1^{'}=0, X_2^{'}=1\right)}\right)-\beta_0^{'}=\log\left(\frac{0.1444}{1-0.1444}\right)+3.1523=1.3731,$$

$$\beta_2^{'}=\log\left(\frac{\Pr\left(p \mid X_1^{'}=1, X_2^{'}=0\right)}{1-\Pr\left(p \mid X_1^{'}=1, X_2^{'}=0\right)}\right)-\beta_0^{'}=\log\left(\frac{0.1444}{1-0.1444}\right)+3.1523=1.3731.$$

So, we can calculate the odds for the genotype (X1' = 1, X2' = 1) based on the logistic regression model as the following,

$$\frac{\Pr\left(p \mid X_1^{'}=1, X_2^{'}=1\right)}{1-\Pr\left(p \mid X_1^{'}=1, X_2^{'}=1\right)}=e^{\beta_0^{'}+\beta_1^{'}\times 1+\beta_1^{'}\times 1}=0.6662$$

However, from Table 7 we know the odd for the genotype (X1' = 1, X2' = 1) is $\frac{0.4266}{1-0.4266} = 0.7440$. So the logistic regression model breaks when tag SNPs are measured.

Example 2.b on measuring tag SNPs of the new model:

In this example, we have two causal SNPs (X1 and X2) that follow the new model,

$$\log(1-p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .$$

Assuming the parameters $\beta_0 = -0.0182$, $\beta_1 = -0.1088$ and $\beta_1 = -0.1088$, we can get Table 8 showing the disease probability under each genotype.

Table 8: Distribution of disease probability with two causal SNPs X1 and X2 under the new model

p	X2 = 0	X2 = 1
X1 = 0	0.0180	0.1192
X1 = 1	0.1192	0. 2100

Assume that we have two measured tag SNPs (X1' and X2') with the same correlation structure as in the example 2.a. With the same calculation procedure we obtain Table 9.

Table 9: Distribution of disease probability with two tag SNPs X1' and X2' under the new model

p	X2' = 0	X2' = 1
X1' = 0	0.0381	0.1183
X1' = 1	0.1183	0.1917

Again, denoting β_0', β_1' and β_2' the parameters for the new model with the tag SNPs, we can compute these parameters based on Table 9 as the following,

$$\beta_0' = \log\left(1 - \Pr\left(p \mid X_1' = 0, X_2' = 0\right)\right) = \log\left(1 - 0.0381\right) = -0.0388,$$

$$\beta_1' = \log\left(1 - \Pr\left(p \mid X_1' = 0, X_2' = 1\right)\right) - \beta_0' = \log\left(1 - 0.1183\right) + 0.0388 = -0.0871,$$

$$\beta_2' = \log\left(1 - \Pr\left(p \mid X_1' = 1, X_2' = 0\right)\right) - \beta_0' = \log\left(1 - 0.1183\right) + 0.0388 = -0.0871.$$

So, we can calculate the odds for the genotype (X1' = 1, X2' = 1) based on the new model as the following,

$$\frac{\Pr\left(p \mid X_1' = 1, X_2' = 1\right)}{1 - \Pr\left(p \mid X_1' = 1, X_2' = 1\right)} = \frac{1 - e^{\beta_0' + \beta_1' \times 1 + \beta_1' \times 1}}{e^{\beta_0' + \beta_1' \times 1 + \beta_1' \times 1}} = 0.2374.$$

On the other hand, from Table 9 we know the odd for the genotype (X1' = 1, X2' = 1) is 0.2374, which gives a numerical confirmation of the theorem of tagged SNPs for the new model.

Example 3.a on the disease heterogeneity of the logistic regression model:

In this example, the disease has three subtypes. We have two SNPs (X1 and X2) under investigation that follow the logistic regression model for each disease subtype,

$$\log\left(\frac{p_n}{1 - p_n}\right) = \beta_{n,0} + \beta_{n,1} x_1 + \beta_{n,2} x_2,$$

where the '*n*' in the subscript denotes the n^{th} subtype, and n =1, 2, or 3.

For the first subtype, assuming the parameters $\beta_{1,0} = -2.5$, $\beta_{1,1} = 1$ and $\beta_{1,2} = 1$, we can get Table 10 showing the disease probability under each genotype.

Table 10: Probability distribution of being the first subtype with genotypes specified by two SNPs X1 and X2 under the logistic regression model.

p_1	X2 = 0	X2 = 1
X1 = 0	0.0754	0.1824
X1 = 1	0.1824	0. 3775

For the second subtype, assuming the parameters $\beta_{2,0} = -3.5$, $\beta_{2,1} = 1.5$ and $\beta_{2,2} = 1.5$, we get Table 11 showing the disease probability under each genotype.

Table 11: Probability distribution of being the second subtype with genotypes specified by two SNPs X1 and X2 under the logistic regression model.

p_2	X2 = 0	X2 = 1
X1 = 0	0.0293	0.1192
X1 = 1	0.1192	0. 3775

For the third subtype, assuming the parameters $\beta_{3,0} = -4.5$, $\beta_{3,1} = 2$ and $\beta_{3,2} = 2$, we get Table 12 showing the disease probability under each genotype.

Table 12: Probability distribution of being the third subtype with genotypes specified by two SNPs X1 and X2 under the logistic regression model.

p_3	X2 = 0	X2 = 1
X1 = 0	0.0110	0.0759
X1 = 1	0.0759	0. 3775

With the assumption of conditional independence among the subtypes of the disease, the distribution of the disease being any of the three subtypes is presented in Table 13 as the following.

Table 13: Probability distribution of being disease (any of the three subtypes) with genotypes specified by two SNPs X1 and X2. Each subtype follows the logistic regression model.

p	X2 = 0	X2 = 1
X1 = 0	0.1123	0.3345
X1 = 1	0.3345	0. 7588

Then, we can get the odds for each genotype as follows,

$$\text{odds}\left(X_1=0, X_2=0\right)=\frac{0.1123}{1-0.1123}=0.1265,$$

$$\text{odds}\left(X_1=0, X_2=1\right)=\frac{0.3345}{1-0.3345}=0.5026,$$

$$\text{odds}\left(X_1=1, X_2=0\right)=\frac{0.3345}{1-0.3345}=0.5026,$$

$$\text{odds}\left(X_1=1, X_2=1\right)=\frac{0.7588}{1-0.7588}=3.1459.$$

However, for the genotype $\left(X_1=1, X_2=1\right)$, based on the other three genotypes and according to the logistic regression model, the odds should be,

$$\text{odds}'\left(X_1=1, X_2=1\right)=\frac{0.5026\times 0.5026}{0.1265}=1.9969.$$

Certainly, $1.9969 \neq 3.1459$. So the logistic regression model is *not* compatible when the disease is heterogeneous.

Example 3.b on the disease heterogeneity of our proposed model:

In this example, the disease has three subtypes. We have two SNPs (X1 and X2) under investigation that follow our proposed model for each disease subtype,

$$\log\left(1-p_n^{'}\right)=\beta_{n,0}^{'}+\beta_{n,1}^{'}x_1+\beta_{n,2}^{'}x_2,$$

where the '*n*' in the subscription denotes the n^{th} subtype, and *n* =1, 2, or 3.

To be comparable to the example provided for the logistic regression model, the parameters for each subtype are set to make the probability distribution the same as for the logistic regression model, except for the genotype $\left(X_1=1,X_2=1\right)$. Please note that it is impossible to make the probability distribution the same for *all* the genotypes.

For the first subtype, assuming the parameters $\beta_{1,0}^{'}=-0.0784$, $\beta_{1,1}^{'}=-0.1230$ and $\beta_{1,2}^{'}=-0.1230$, we can get Table 14 showing the disease probability under each genotype.

Table 14: Probability distribution of being the first subtype with genotypes specified by two SNPs X1 and X2 under the new model.

p_1	X2 = 0	X2 = 1
X1 = 0	0.0754	0.1824
X1 = 1	0.1824	0. 2774

For the second subtype, assuming the parameters $\beta_{2,0}^{'}=-0.0297$, $\beta_{2,1}^{'}=-0.0972$ and $\beta_{2,2}^{'}=-0.0972$, we can get Table 15 showing the disease probability under each genotype.

Table 15: Probability distribution of being the second subtype with genotypes specified by two SNPs X1 and X2 under the new model.

p_2	X2 = 0	X2 = 1
X1 = 0	0.0293	0.1192
X1 = 1	0.1192	0. 2008

For the third subtype, assuming the parameters $\beta_{3,0}^{'} = -0.0110$, $\beta_{3,1}^{'} = -0.0679$ and $\beta_{3,2}^{'} = -0.0679$, we can get Table 16 showing the disease probability under each genotype.

Table 16: Probability distribution of being the third subtype with genotypes specified by two SNPs X1 and X2 under the new model.

p_3	X2 = 0	X2 = 1
X1 = 0	0.0110	0.0759
X1 = 1	0.0759	0. 1366

With the assumption of conditional independence among the subtypes of the disease, the distribution of the disease being any of the three subtypes is presented in Table 17 as the following.

Table 17: Probability distribution of being disease (any of the three subtypes) with genotypes specified by two SNPs X1 and X2. Each subtype follows the new model.

p	X2 = 0	X2 = 1
X1 = 0	0.1123	0.3345
X1 = 1	0.3345	0. 5011

Then, we can get the odds for each genotype as follows,

$$\text{odds}\left(X_1 = 0, X_2 = 0\right) = \frac{0.1123}{1-0.1123} = 0.1265,$$

$$\text{odds}\left(X_1 = 0, X_2 = 1\right) = \frac{0.3345}{1-0.3345} = 0.5026,$$

$$\text{odds}\left(X_1 = 1, X_2 = 0\right) = \frac{0.3345}{1-0.3345} = 0.5026,$$

$$\text{odds}\left(X_1 = 1, X_2 = 1\right) = \frac{0.5011}{1-0.5011} = 1.0044.$$

For the genotype $(X_1=1, X_2=1)$, based on the other three genotypes and according to our proposed model, we have that,

$$1-\Pr(dis \mid X_1=1, X_2=1)=\frac{(1-0.3345)\times(1-0.3345)}{1-0.1123}=0.4989$$

$$\text{odds}'(X_1=1, X_2=1)=\frac{1-0.4989}{0.4989}=1.0044.$$

So our proposed model is indeed compatible when the disease is heterogeneous. Specifically, we can estimate the parameters of our proposed model based on Table 17, such as, $\beta_0'=\log(1-0.1123)=-0.1191$, $\beta_1'=\log(1-0.3345)-\beta_0'=-0.2881$ and $\beta_2'=-0.2881$. Comparing to the parameters for each subtype, we get that $\beta_0'=\beta_{1,0}'+\beta_{2,0}'+\beta_{3,0}'$, $\beta_1'=\beta_{1,1}'+\beta_{2,1}'+\beta_{3,1}'$ and $\beta_2'=\beta_{1,2}'+\beta_{2,2}'+\beta_{3,2}'$, which gives a numerical confirmation of the theorem of disease heterogeneity for the new model.

II.1.2.E Power of the proposed statistic compared to logistic regression with interaction terms

Type I error rate and power are two key aspects evaluating a hypothesis testing approach. As we have shown previously, the null hypothesis of logistic regression is generally violated even if there is indeed no interaction effect. The logistic regression model does not hold for many common situations such as the three we mentioned above, so it can be easily violated. The violation induces an inaccurate significance assessment.

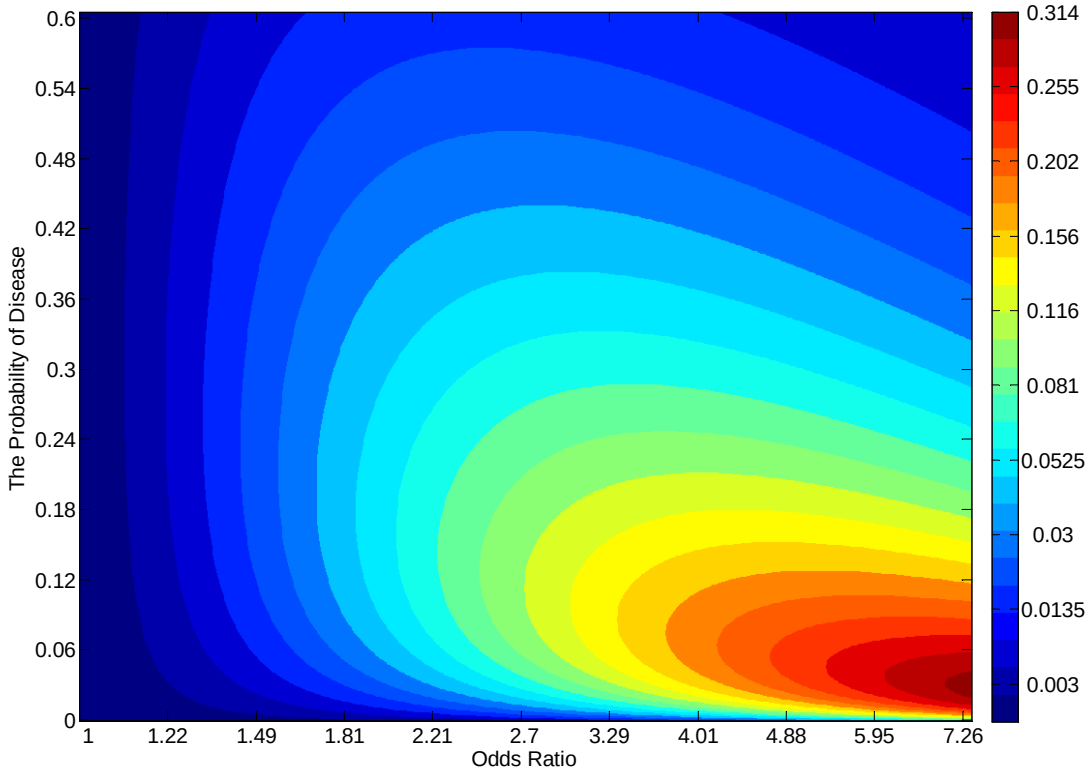

Figure 5: The difference of predicted probability of disease by logistic regression model and our model. The larger the difference is, the more the power is reduced for logistic regression model.

Then, how is the power of the logistic regression model, compared to our new proposed model? We have mathematically proved that the new model is strictly more powerful than the logistic regression model as shown in ***Theorem 8*** and ***Corollary 8.1***. The theoretical analysis demonstrates that, as long as the factors are genuine disease-risk factors (possessing some marginal effects) and have true interaction effects, the new model will generate a smaller p-value than the logistic regression model on *any* dataset. Figure 5 shows the difference of predicted probability of disease by the logistic regression model and by our model. The larger the difference is, the more the power is reduced for logistic regression model. We can see from the figure that the difference is always positive, which is equivalent to say the new model will always have better power to detect interaction. The increase of power depends on the parameters

such as the odds ratio and the fraction of cases in study, and the dependence is nonlinear. In general, the new model will have more power gain for large effect size (large odds ratio) and less fraction of cases (but not extremely less).

II.1.2.F Brief summary and intuitive explanations on identifying interactions among significant disease factors

The flow chart for identifying interactions among significant disease factors is shown in Figure 1.c. The identification of interactions among known disease-related factors is derived from the framework of the likelihood ratio test. The most critical parts are the choice of the probability model to fit to the problem and hence the induced optimization problem. By assuming the linearity of logarithm of the probability of being *healthy*, we have a model that possesses three critical properties desirable in practice and proves to have better power than conventional logistic regression.

Symmetry is the key difference between the conventional logistic regression model and the new model. The logistic regression model is symmetric. When we reverse the label of case and control, the likelihood is kept the same for the logistic regression model. The new model is asymmetric and the likelihood changes under label reversal. At the same time, the disease label is asymmetric, because, for one individual to be free of disease each of the disease risk factors that is present must not cause the disease. Likewise, an individual gets the disease so long as one disease risk factor penetrates.

In the process of devising the approach of identifying interactions among significant disease factors, we made the follow mathematical contributions. (1) We have figured out several important properties to be satisfied for a criterion of identifying the interaction with significant marginal effect to be applicable in practical situation, which, include the compatibility with hidden risk factor, tag markers and disease heterogeneity, and, are mathematically verified to be true for our proposed criterion. (2) We have proved theoretically that our approach is more powerful than the conventional logistic regression with interaction terms in a strict sense, that is, if the interaction is true, the new model will generate a smaller p-value on any dataset.

Lemma 5.1: **Denote** $\Pr(dis \mid \mathrm{X}) = \Pr(\mathrm{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_M = x_M)$ **and** $\Pr(dis \mid \mathrm{X}_{-i}) = \Pr(\mathrm{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_{i-1} = x_{i-1}, \mathrm{X}_{i+1} = x_{i+1}, ..., \mathrm{X}_M = x_M)$ **. Assume that** $\Pr(\mathrm{X}_i = x_i \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_{i-1} = x_{i-1}, \mathrm{X}_{i+1} = x_{i+1}, ..., \mathrm{X}_M = x_M) = \Pr(\mathrm{X}_i = x_i)$ **. If** $\log(1 - \Pr(dis \mid \mathrm{X})) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$ **, then** $\log(1 - \Pr(dis \mid \mathrm{X}_{-i})) = \beta_0' + \sum_{m=1, m \neq i}^{M} \beta_m x_m$ **, where** $\beta_0' = \beta_0 + \log\left(\sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$

Theorem 5: **Denote** $\Pr(dis \mid \mathrm{X}) = \Pr(\mathrm{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_M = x_M)$**. Assume that** $\mathrm{X}_1, ..., \mathrm{X}_M$ **are statistically independent. If** $\log(1 - \Pr(dis \mid \mathrm{X})) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$ **, then for any subset of** $\mathrm{S} \subset \{\mathrm{X}_1, ..., \mathrm{X}_M\}$ **, we have** $\log(1 - \Pr(dis \mid \mathrm{S})) = \beta_0' + \sum_{X_m \in S} \beta_m x_m$ **, where** $\beta_0' = \beta_0 + \sum_{X_i \in \bar{S}} \log\left(\sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$**.**

Lemma 6.1: **Denote** $\Pr(dis \mid \mathrm{X}) = \Pr(\text{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_M = x_M)$ **. Assume** X_i **is not measured and a correlated variable** $\mathrm{X}_i^{'}$ **is observed. Assume the conditional independence relation that** $\Pr(\mathrm{X}_i \mid \mathrm{X}_1, ..., \mathrm{X}_{i-1}, \mathrm{X}_i^{'}, \mathrm{X}_{i+1}, ..., \mathrm{X}_M) = \Pr(\mathrm{X}_i \mid \mathrm{X}_i^{'})$**. Write** $\Pr(\mathrm{X}_i = x_i | \mathrm{X}_i^{'} = x_i^{'}) = p(x_i^{'}, x_i)$**. Denote** Ω_m **the range of** X_m **and** ν_m **the smallest element in** Ω_m**,** $m = 1, ..., M$ **. Denote** $\Omega_i^{'}$ **the range of** $\mathrm{X}_i^{'}$ **and** $\nu_i^{'}$ **the smallest element in** $\Omega_i^{'}$**. If** $\log(1 - \Pr(dis \mid \mathrm{X})) = \beta_0 + \sum_{m=1}^{M} \left[\sum_{\omega_m \in \Omega_m \setminus \nu_m} \beta_m(\omega_m) I(\omega_m = x_m) \right]$**, then we have** $\log(1 - \Pr(dis \mid \mathrm{X}_1, ..., \mathrm{X}_{i-1}, \mathrm{X}_i^{'}, \mathrm{X}_{i+1}, ..., \mathrm{X}_M)) = \beta_0^{'} + \sum_{m \neq i} \left[\sum_{\omega_m \in \Omega_m \setminus \nu_m} \beta_m(\omega_m) I(\omega_m = x_m) \right] + \sum_{\omega_i^{'} \in \Omega_i^{'} \setminus \nu_i^{'}} \beta_i^{'}(\omega_i^{'}) I(\omega_i^{'} = x_i^{'})$**.**

Theorem 6: X_m**,** $m = 1, ..., M$ **are** M **disease-causing markers but not directly measured.** $\mathrm{X}_m^{'}$**,** $m = 1, ..., M$ **are M observed markers correlated to** X_m**, respectively. Assume the conditional independence that** $\Pr(\mathrm{X}_i \mid \mathrm{X}_1, ..., \mathrm{X}_{i-1}, \mathrm{X}_i^{'}, \mathrm{X}_{i+1}, ..., \mathrm{X}_M) = \Pr(\mathrm{X}_i \mid \mathrm{X}_i^{'})$**. Denote** Ω_m **the range of** X_m **and** ν_m **the smallest element in** Ω_m**,** $m = 1, ..., M$ **. Denote** $\Omega_m^{'}$ **the range of** $\mathrm{X}_m^{'}$ **and** $\nu_m^{'}$ **the smallest element in** $\Omega_m^{'}$**,** $m = 1, ..., M$ **. If** $\log(1 - \Pr(dis \mid \mathrm{X})) = \beta_0 + \sum_{m=1}^{M} \left[\sum_{\omega_m \in \Omega_m \setminus \nu_m} \beta_m(\omega_m) I(\omega_m = x_m) \right]$**, then** $\log(1 - \Pr(dis \mid \mathrm{X}))$ **follows a similar equation and** $\log(1 - \Pr(dis \mid \mathrm{X}^{'})) = \beta_0^{'} + \sum_{m=1}^{M} \left[\sum_{\omega_m^{'} \in \Omega_m^{'} \setminus \nu_m^{'}} \beta_m^{'}(\omega_m^{'}) I(\omega_m^{'} = x_m^{'}) \right]$**.**

Corollary 6.1: X_m**,** $m = 1, ..., M$ **are** M **disease-causing markers but not directly measured.** $\mathrm{X}_m^{'}$**,** $m = 1, ..., M$ **are M observed markers correlated to** X_m**, respectively. Assume the conditional independence that** $\Pr(\mathrm{X}_i \mid \mathrm{X}_1, ..., \mathrm{X}_{i-1}, \mathrm{X}_i^{'}, \mathrm{X}_{i+1}, ..., \mathrm{X}_M) = \Pr(\mathrm{X}_i \mid \mathrm{X}_i^{'})$**. Denote** Ω_m **the range of** X_m **and** ν_m **the smallest element in** Ω_m**,** $m = 1, ..., M$ **. If** $\mathrm{X}_m^{'}$ **are all binary with only two possible values** $\{0,1\}$

and $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\left[\sum_{\omega_m\in\Omega_m\setminus\nu_m}\beta_m\left(\omega_m\right)I\left(\omega_m=x_m\right)\right]$ **, we have,** $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}^{'}\right)\right)=\beta_0^{'}+\sum_{m=1}^{M}\beta_m^{'}x_m^{'}$ **.**

Corollary 6.2: X_m **,** $m=1,...,M$ **are** M **disease-causing markers but not directly measured.** $\mathrm{X}_m^{'}$ **,** $m=1,...,M$ **are M observed markers correlated to** X_m **, respectively. Assume the conditional independence that** $\Pr\left(\mathrm{X}_i\,|\,\mathrm{X}_1,...,\mathrm{X}_{i-1},\mathrm{X}_i^{'},\mathrm{X}_{i+1},...,\mathrm{X}_M\right)=\Pr\left(\mathrm{X}_i\,|\,\mathrm{X}_i^{'}\right)$ **. If** $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\left[\sum_{\omega_m\in\Omega_m\setminus\nu_m}\beta_m\left(\omega_m\right)I\left(\omega_m=x_m\right)\right]$ **, after dominant/recessive genetic coding of** $\mathrm{X}_m^{'}$ **, we have,** $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}^{'}\right)\right)=\beta_0^{'}+\sum_{m=1}^{M}\beta_m^{'}x_m^{'}$ **.**

Theorem 7: **If the given disease '** dis **' contains** N **subtypes, that is,** $dis=\bigcup_{n=1}^{N}dis_n$ **. Denote X the disease factors under investigation and** $\Pr\left(dis_n\,|\,\mathrm{X}\right)=\Pr\left(\mathrm{Status}=dis_n\,|\,\mathrm{X}_1=x_1,...,\mathrm{X}_M=x_M\right)$ **. Assume the conditional independence among** dis_n **, that is,** $\Pr\left(dis_1,...,dis_N\,|\,\mathrm{X}\right)=\prod_{n=1}^{N}\Pr\left(dis_n\,|\,\mathrm{X}\right)$ **. If** $\log\left(1-\Pr\left(dis_n\,|\,\mathrm{X}\right)\right)=\beta_{n,0}+\sum_{m=1}^{M}\beta_{n,m}x_m$ **for** $n=1,...,N$ **, then we have** $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$ **, where** β_0 **and** β_m **are constants independent to X, and more specifically,** $\beta_0=\sum_{n=1}^{N}\beta_{n,0}$ **and** $\beta_m=\sum_{n=1}^{N}\beta_{n,m}$ **.**

Theorem 8: X_1 **and** X_2 **are two binary variables. Suppose we know** $\Pr\left(dis\,|\,\mathrm{X}_1=0,\mathrm{X}_2=0\right)=p_{00}$ **,** $\Pr\left(dis\,|\,\mathrm{X}_1=0,\mathrm{X}_2=1\right)=p_{01}$ **and** $\Pr\left(dis\,|\,\mathrm{X}_1=1,\mathrm{X}_2=0\right)=p_{10}$ **,** $p_{01}\geq p_{00}$ **and** $p_{10}\geq p_{00}$ **. Denote** p_{11} **the predicted value of** $\Pr\left(dis\,|\,\mathrm{X}_1=1,\mathrm{X}_2=1\right)$ **according to the logistic regression model** $\log\left(\Pr\left(dis\,|\,\mathrm{X}\right)/\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$ **. Denote** $p_{11}^{'}$ **the predicted value of**

$\Pr(dis \mid X_1 = 1, X_2 = 1)$ **according to the new model** $\log(1 - \Pr(dis \mid X)) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$ **. We have** $p_{11} \geq p_{11}^{'}$**, where the equality holds if and only if** $p_{01} = p_{00}$ **or** $p_{10} = p_{00}$**.**

Corollary 8.1: $X_1 \in \{0,1\}$ **and** $X_2 \in \{0,1\}$ **are two binary variables with '1' denoting the disease-risk increasing genotypes. Suppose** X_1 **and** X_2 **have interactions and that** $\Pr(dis \mid X_1 = 1, X_2 = 1)$ **is larger than the value predicted by the logistic regression model. Then, the detection power through logistic regression is smaller than the power through the model** $\log(1 - \Pr(dis \mid X)) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$ **.**

II.1.3 Efficient Heuristic Search Algorithm and Other Computational Techniques

When the interaction effects come in, the computation and search in GWAS studies become a daunting challenge because the combination of SNPs subsets is a huge number. Instead of exhaustively searching all the possibilities, people try to reduce the computation by focusing on the more possible candidates, such as greedy search or screening the SNPs into a small pool by examining the marginal p-values. As discussed later, these methods will miss a lot of true targets. More seriously, the existing heuristic searching algorithms involve a lot of parameters controlling the accuracy and computation burden, while the choice of these parameters is more like to be an art with no guidance on what you expect to get. With the introduction of two new concepts, the *transferable genotype-effect potential* (TGEP) and the *worst situation*, we propose a novel searching scheme efficient both theoretically and practically with guaranteed lower-bound of performance. We provide the user the control on the performance at the most difficult situation. Note here that the searching algorithm pertains only to the detection of informative SNPs with the interaction effects incorporated. We have used the exhaustive search to identify

the interaction among SNPs with significant marginal effects, because the number of SNPs with significant marginal effects is small compared to the original SNPs genotyped, usually at the scale of less than tens.

The proposed heuristic search has explored and harnessed both the inherited properties of the GWAS data and the constructed properties from our new concepts and algorithm design, such as the invariant property of TGEP, the relaxation of the most difficult situation, the finiteness of p-value generated from the hypergeometric distribution, the low p-value threshold to achieve GWAS significance, and the non-significance of lower-order subsets from the re-utilization of existing discoveries. In addition, besides the fast implementation of heypergeometric cumulative function discussed previously, we have also applied various techniques to speed the computation, for instance, using the hash table to reduce the redundant search, adopting the direct table to reduce the repetitive computation in calculating the combination number, and employing the hierarchical structure to take advantage of partial-duplicated counting.

II.1.3.A Definition of the transferable genotype-effect potential (TGEP)

Suppose there are totally *M* subjects, of which *j* are cases and *M-j* are controls. g is a disease-risk genotype specified by multiple SNPs. The genotype g specifies a population of size r with k cases and $r-k$ controls. The complex disease may have many disease factors other than the genotype g, which means, the *k* cases may be caused by the genotype *g*, the other disease factors, or both. We are interested in the phenotypic subject (here, the phenotype stands for the case) induced exclusively by the genotype *g*. We name this type of subjects as the transferable genotype-effect potential (TGEP), that is, TGEP is induced only by the specific genotype, not by

any other factor. Actually, unless under some extreme scenarios such as that the genotype and other factors never occur simultaneously or with some other information like further biological experiments, we are not able to decide whether a phenotypic subject is TGEP. However, given a disease-risk genotype *g*, we could estimate the expected number Υ of TGEP as the following,

$$\Upsilon = k - \frac{j-k}{M-r} r \tag{24}$$

Intuitively, the number Υ can be estimated by subtracting the expected cases induced by all the other factors from the *k* cases. $(j-k)/(M-r)$ can be seen as the estimation of the rate of cases induced by all other factors. Timing this rate with the number *r* of subjects carrying the genotype g, we obtain the expected number of cases with disease causes from other factors and accordingly the number Υ.

One important attribute of TGEP is the invariant property as proved in ***Theorem 9*** and ***Corollary 9.1.*** Basically, the invariant property says that the expectation remains invariant if the number Υ is estimated through a population larger than the one specified by the genotype g. Specifically, the number Υ associated with a high order genotype can be estimated unbiasedly through its lower order genotype. For example, to obtain the number Υ associated with a three-order genotype $[s_1 = 0, s_2 = 1, s_3 = 1]$, we can estimate it through the lower-order genotypes which have totally six possible types as, $[s_1 = 0, s_2 = 1]$, $[s_1 = 0, s_3 = 1]$, $[s_2 = 1, s_3 = 1]$, $[s_1 = 0]$, $[s_2 = 1]$ and $[s_3 = 1]$. The invariant property has important implications in the design of heuristic search. It enables us to deal with different-order interactions systematically and gather information from the low-order SNP subset to decode the high-order interaction.

The invariant property of TGEP is illustrated by the example in the Figure 6. The TGEP is associated with the genotype (X1=1, X2=1). There are three possible ways to calculate the number of TGEP, (1) from the statistics of X1 only, (2) from the statistics of X2 only, and (3) from the statistics of both X1 and X2. From the example, we can see that the three ways generate the same estimation of the number of TGEP, which indicates the plausibility of inferring TGEP associated with the higher-order interaction from its lower-order (marginal) statistics.

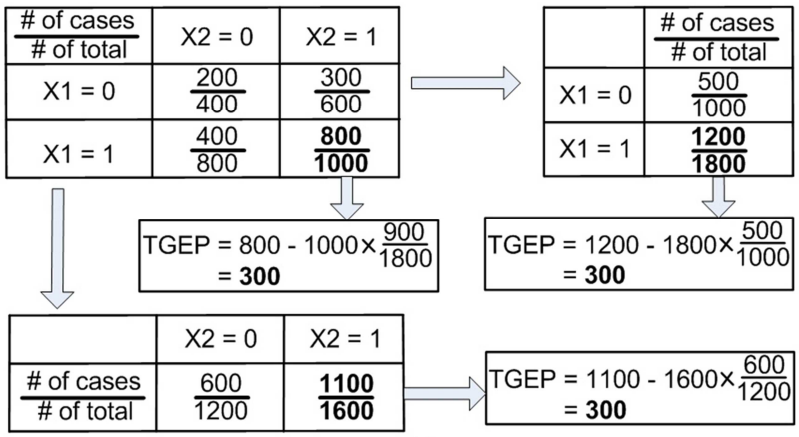

Figure 6: Illustration of the invariant property of TGEP.

The number Υ has a direct link to the significance of SNP subset. Denote α the frequency of cases. Denote B the frequency of the disease-risk genotype g. Then, we have $\alpha = j/M$ and $B = r/M$. Through simple algebra operations, we can rewrite the Υ as,

$$\Upsilon = k - \frac{j-k}{M-r} r = \frac{Mk - jr}{M-r} = \frac{k - jB}{1-B} \qquad (25)$$

From the hypergeometric distribution, we know that k approximately follows the Gaussian distribution with mean jB and variance $M\alpha(1-\alpha)B(1-B)$. Thus, denoting $\Phi(\bullet)$ as the cumulative distribution function of standard Gaussian variable, the right-tail p-value associated with the corresponding SNP subset is,

$$p \approx \Phi\left(-\frac{k-jB}{\sqrt{M\alpha(1-\alpha)B(1-B)}}\right) = \Phi\left(-\frac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B}-1}\right) \quad (26)$$

For the sake of discussion, we let,

$$z = \frac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B}-1} \quad (27)$$

Because M and α are fixed for a certain experiment, the z score depends on Υ and B. From the property that Υ remains invariant with different-order genotypes, we could estimate Υ based on the lower-order genotype data. In addition, we could also estimate the frequency B of the high-order genotype from its lower-order data. Thus, it is possible to estimate the significance of high-order SNP subset from its lower-order information. The detailed formulation of the estimation is described in the subsection II.1.3.C.

One by-product of the definition of TGEP is to illustrate how the incorporation of interaction effects potentially improves the detection power of informative SNPs. Because of the invariant property of Υ across genotypes with different orders, from the equation (27) we can see that the significance purely relies on the frequency of the genotype under investigation and increases rapidly with B decreasing. To make clear of this point, let us consider a three-way interaction with each risk-allele for each SNP having the frequency 0.5. Here, we assume the SNP only takes two values, risk and non-risk. Thus, the three-SNP disease-risk genotype will have a frequency B of 0.125. Readily having that $\sqrt{1/0.125-1} = 2.65\times\sqrt{1/0.5-1}$, the z score after incorporating the interaction is 2.65 times of the z score based on the marginal effect. Recalling the p-value decreases super-exponentially with the z score, the 2.65 fold will increase the significance significantly. Hence, the three-way interacting SNP subset with a p-value of 1×10^{-6}

for the marginal effect (which is not significant for 1 million tests) could result in a p-value of 1.1×10^{-36} by incorporating the interaction effects.

The definition of TGEP also shows the pre-screening based on the marginal p-values and greedy search are theoretically sub-optimum. The marginal p-value is a combined effect of both the number Υ of TGEP and the frequency *B* of risk-allele. A large marginal p-value may come from a small Υ or a large *B*. It is possible that a large Υ shows very small marginal effects. On the other hand, the greedy search relies heavily on the previous selected SNPs, while the interacting SNPs may show no significant marginal effects as discussed above and thus fail the greedy search.

Theorem 9: **The expected number of TGEP is an invariant quantity. Assume g_1 and g_2 are two mutually disjoint genotypes, of which g_1 a disease-risk genotype is and g_2 is not a disease-risk genotype. Denote $\Upsilon(g_1)$ and $\Upsilon(g_1 \cup g_2)$ the estimated numbers based on g_1 and the union of g_1 and g_2, respectively. Then, $E(\Upsilon(g_1)) = E(\Upsilon(g_1 \cup g_2))$.**

Corollary 9.1: **Assume the genotype *g* is the only disease-risk genotype specified by SNPs subset S containing *n* interacting SNPs, the number of TGEP associated with *g* can be estimated by its lower order genotype specified by a proper subset of S.**

II.1.3.B Definition of the worst situation

The worst situation in our heuristic search scheme is the case that has the following two characteristics: (1) the p-value of the interacting SNP subset is at **the borderline** between

experimental-wise significant and insignificant, and (2) the risk allele is such **rare** that it is the minimum capable frequency, that is, a smaller allele frequency will not be capable to generate an experimental-wise significant p-value due to the finiteness of p-value associated with hyper-geometric distribution (***Proposition 1***). The worst situation is defined in the way such that any significant interacting SNP subset has no worse detection rate than the worst situation.

We are not interested in the interacting SNP subset with p-value larger than the threshold of experimental-wise significance, because the data has no power to detect the SNP subset whatever it is searched or not. As we know, the more significant the interacting SNP subset is, the larger effects its lower-order subsets have. Thus, the worst case requires that the p-value is at the borderline. On the other hand, when we fix the p-value, the smaller the frequency of the disease-risk genotype, the smaller effects its lower-order subsets have, since the effect of the disease-risk genotype will be diluted at lower-order level. Hence, the worst situation also requires the frequency of the disease-risk genotype as small as possible. However, the frequency cannot be infinitely small since the hypergeometric distribution has a non-zero minimum p-value as shown in ***Proposition 1*** and the frequency has to be large enough to be able to generate a borderline significant p-value.

Denote M as the number of total subjects and j as the number of cases. Let P_t be the threshold of experimental-wise significance. Write $\alpha = j/M$. Let $r_t = \Gamma^{-1}(P_t)$, with $\Gamma(r)$ being the minimum p-value as defined in ***Theorem 2*** of a hypergeometric distribution of parameter set $[M, j, r]$. At the worst situation, we have the formulas for the frequency B of the disease-risk genotype and the number Υ of TGEP as the following,

$$B_t = \frac{r_t}{M} \tag{28}$$

$$\Upsilon_t = \frac{1-\alpha}{1-B_t} r_t \tag{29}$$

The equations (28) and (29) are the necessary and sufficient conditions for the worst situation.

The worst situation is the extremely difficult case and it should rarely occur in real application. There are two possible ways for a significant SNP subset to deviate from the worst situation and thus make it easier to detect the informative SNPs. One way is to increase the frequency of the disease-risk genotype but with the p-value remaining at P_t. The other way is to decrease the p-value. In either way, we will have $\Upsilon > \Upsilon_t$, which can be proved by utilizing the equations (26) and (27). Let B and Υ present the new frequency and number of TGEP, respectively. In the first scenario, we have $B > B_t$ and the p-values unchanged, so we get,

$$\frac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B}-1} = \frac{\Upsilon_t}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B_t}-1}$$

$$\Rightarrow \Upsilon\sqrt{\frac{1}{B}-1} = \Upsilon_t\sqrt{\frac{1}{B_t}-1}$$

$$\Rightarrow \Upsilon > \Upsilon_t \qquad (\because B > B_t) \tag{30}$$

In the second scenario, because the p-value is decreased, we have,

$$\frac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B}-1} > \frac{\Upsilon_t}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B_t}-1} \tag{31}$$

Since B_t is the minimum frequency to generate a p-value as small as P_t, B has to be larger than B_t, otherwise B_t is not the minimum capable frequency. Thus, reforming the equation (31) and taking the fact $B > B_t$, we have,

$$\Upsilon\sqrt{\frac{1}{B}-1} > \Upsilon_t\sqrt{\frac{1}{B_t}-1}$$

$$\Rightarrow \frac{\Upsilon}{\Upsilon_t} > \sqrt{\frac{1}{B_t}-1}\Big/\sqrt{\frac{1}{B}-1} > 1 \quad (\because B > B_t) \qquad (32)$$

From the above proof, we know a SNP subset with $\Upsilon < \Upsilon_t$ is impossible to be experimental-wise significant. Hence, it will not lose any power if we avoid searching the SNP subsets with $\Upsilon < \Upsilon_t$. This also suggests an approach of relaxing the worst situation to improve the computation speed by discarding the SNP subsets with $\Upsilon < \Upsilon_t^{'}$, where $\Upsilon_t^{'} > \Upsilon_t$. Here, the invariant property of Υ plays an important role, because we need to check Υ before we really interrogate the SNP subset and the invariant property makes us able to check Υ based on its lower-order information. Actually, we can see from the above proof that $B \geq B_t$ for any significant SNP subset. But B does not have the nice invariant property and cannot be utilized to direct the heuristic search. The relaxation of the worst situation is not only probabilistically sound as the probability of the occurrence of the worst situation will be small, but also it is supported by the biological sense especially in terms of common disease. With the common disease, it is hardly to expect that the disease-risk genotype will have a 100% penetrance, while the worst situation corresponds to the 100% penetrance. So, it is pretty safe to relax the worst situation.

II.1.3.C Heuristic combinatorial interaction growing algorithm (HCIG)

We propose a heuristic search algorithm based on the seeds growing. To detect the d^{th} order interaction, we sequentially grow the first order seeds, the second order seeds, to the $(d\text{-}1)^{th}$ order seeds. Then, candidates for the d^{th} order interaction are constructed by pairing the $(d\text{-}1)^{th}$ order seeds and the rest of SNPs. Two heuristic schemes working together are employed to guide the search towards the SNP combinations with the most potential to be real interactions. The first scheme is created for the growing of the seeds and the second scheme is designed for the construction of the final candidates for the d^{th} order interaction.

Strategy of seeds growing

Here, we show how to generate an $(i+1)^{th}$ order seed from the i^{th} order seed. Suppose $S^i = \{s_1, ...s_i\}$ is the i^{th} order seed containing i SNPs and s is a SNP not included in S^i. The seed S^i specifies a genotype g_i indicating the potential interaction. Denote M_0 as the number of total subjects and j_0 as the number of cases. Write $\alpha_0 = j_0 / M_0$ as the ratio of cases. Denote M_i as the number of subjects carrying the genotype g_i and j_i as the number of cases carrying the genotype g_i. Write $\alpha_i = j_i / M_i$. Let P_d and P_{d-1} be the thresholds of experimental-wise significance for the d-order and $(d\text{-}1)$ order interactions, respectively. With $\Gamma(\bullet)$ defined as in ***Theorem 2,*** let $r_d = \Gamma^{-1}(P_d)$ as the minimum size to generate a p-value as significant as P_d. Write $B_d = r_d / M_i$. With g denoting a genotype specified by the SNP s, denote r as the number of subjects carrying the genotype $g_i \cap g$, of which k is the number of cases. Let us denote $B = r / M_i$. According to the equation (24) and recalling the invariant property, based on the

population conditioned on the genotype g_i the number of TGEP associated with $g_i \cap g$ can be estimated as,

$$\Upsilon = k - \frac{j_i - k}{M_i - r} r \tag{33}$$

Then, for $S^i \bigcup \{s\}$ to be an $(i+1)^{\text{th}}$ order seed, it requires that Υ satisfies the following conditions,

$$\begin{cases} \Upsilon \geq r_d \left(1 + B_d - \alpha_i\right) + c_{i+1}\sigma & \text{if } B_{\min} < B_d \\ \Upsilon \geq Z_d \sqrt{\dfrac{M_0 \alpha_0 \left(1-\alpha_0\right) B_{\min}}{1 - B_{\min}}} + c_{i+1}\sigma & \text{if } B_{\min} \geq B_d \end{cases} \tag{34}$$

where,

c_{i+1} is a parameter to control the likelihood of missing the real interaction,

$\sigma^2 = \dfrac{M_i \alpha_i \left(1-\alpha_i\right) B}{1-B}$,

$F = \left(\dfrac{\Phi^{-1}\left(P_d\right)}{\Phi^{-1}\left(P_{d-1}\right)} \right)^2$, (here, $\Phi(\bullet)$ is the CDF of standard normal distribution)

$Z_d = -\Phi^{-1}\left(P_d\right)$,

and $B_{\min} = \dfrac{BF - 1}{F - 1}$.

Construction of candidates for the d-order interaction from the (d-1)th order seeds

Now, we describe how to construct the candidates for the *d*-order interaction detection by pairing the $(d\text{-}1)^{\text{th}}$ order seed and the single SNP in the rest SNP pool. Suppose $S^{d-1} = \{s_1, ...s_{d-1}\}$ is the $(d\text{-}1)^{\text{th}}$ order seed containing (*d*-1) SNPs and s is a SNP not included in S^d. Denote g_1 and g_2 as genotypes specified by S^{d-1} and s, respectively. Suppose there are totally M subjects with j

subjects being cases. The genotype g_1 specifies a population with r_1 subjects and k_1 cases. The genotype g_2 specifies a population with r_2 subjects and k_2 cases. Let Υ_1 denote the number of TGEP associated with $g_1 \bigcap g_2$ estimated through the genotype g_1 and Υ_2 denote the number of TGEP associated with $g_1 \bigcap g_2$ estimated through the genotype g_2. We have,

$$\Upsilon_1 = k_1 - \frac{j - k_1}{M - r_1} r_1 \tag{35}$$

$$\Upsilon_2 = k_2 - \frac{j - k_2}{M - r_2} r_2 \tag{36}$$

Then, according to ***Theorem 10***, the best estimation of the number of TGEP associated with $g_1 \bigcap g_2$ by combining Υ_1 and Υ_2 and the variance of the estimation are,

$$\Upsilon = \frac{\sigma_2^2 \Upsilon_1 + \sigma_1^2 \Upsilon_2}{\sigma_2^2 + \sigma_1^2} \tag{37}$$

$$\sigma^2 = \mathrm{var}(\Upsilon) = \frac{\sigma_2^2 \sigma_1^2}{\sigma_2^2 + \sigma_1^2} \tag{38}$$

where,

$$\sigma_1^2 = \frac{M(1 - B_1 B_2)\alpha_1(1 - \alpha_1) B_1^{'}(1 - B_1^{'})}{(1 - B_1)^2}$$

$$\sigma_2^2 = \frac{M(1 - B_1 B_2)\alpha_2(1 - \alpha_2) B_2^{'}(1 - B_2^{'})}{(1 - B_2)^2}$$

$$\alpha_1 = \alpha - \frac{\Upsilon_1}{M}$$

$$\alpha_2 = \alpha - \frac{\Upsilon_2}{M}$$

$$B_1^{'} = \frac{B_1(1-B_2)}{1-B_1B_2}$$

$$B_2^{'} = \frac{B_2(1-B_1)}{1-B_1B_2}$$

Denote P_d the threshold of the experimental-wise significance for the *d*-order interaction. Then, for $S^{d-1} \cup \{s\}$ to be a candidate SNP subset for the *d*-order interaction, it requires that Υ satisfies the following condition,

$$\Upsilon \geq Z_d \sqrt{\frac{M\alpha(1-\alpha)B_1B_2}{1-B_1B_2}} + c_d\sigma \qquad (39)$$

where c_d is a control parameter and $Z_d = -\Phi^{-1}(P_d)$.

Setting the control parameters

There are *d* parameters $(c_i, i = 1,...,d)$ left undetermined in the equations (34) and (39). Let τ denote the detection rate at the worst situation (the lower bound of the performance). Denote $\Phi^{-1}(\bullet)$ the inverse CDF of standard Gaussian distribution. Given a value *h* $(0 \leq h \leq 1)$, with the intention of controlling that $\tau \geq h$, the parameters can be set as the following,

$$\begin{cases} c_d = -\Phi^{-1}(v_d) \\ c_i = \left(\sqrt{2}\right)^{i-1} f^{-1}(h/v_d) \qquad i = 1,...,d-1 \end{cases} \qquad (40)$$

where $v_d = \max(0.95, 0.5 + h/2)$ and $f(x) = \prod_{i=1}^{d-1}\left(1 - \left(\Phi\left(\left(\sqrt{2}\right)^{i-1} x\right)\right)^{d-i+1}\right)$. For $d > 2$, there is no close form for $f^{-1}(\bullet)$. Fortunately, from ***Theorem 11*** we know that $f(x)$ is a strictly

decreasing function and there exists an efficient algorithm with logarithm complexity to solve $f^{-1}(\bullet)$.

Analysis of the performance of HCIG

Given a value h $(0 \le h \le 1)$, with the searching schemes as in the equations (34) and (39) and the setting of the parameters as in the equation (40), any significant interaction will have the detection rate larger than *h*. Denote τ the detection rate at the worst situation. Now, we first calculate τ and show $\tau \ge h$, then we illustrate the detection rate for any significant interaction is larger than τ.

Denote ζ_i the survival rate for the significant interaction passing the *i*-order screening, where *i* is from 1 to *d*. The first $(d-1)$ survival rates are for the seeds growing and the last one is for the construction of the final interaction candidate. According to ***Theorem 12***, there is a prerequisite that the frequency of the disease risk genotype g_d is larger than $B_{\min} = (BF-1)/(F-1)$. On the other hand, from the property of the worst situation we know that $B_{\min} \ge B_d$ as in the equation (34). Denote Υ_0 the number of TGEP at the worst situation. Thus, at the worst situation, we have,

$$\begin{cases} \Upsilon_0 = r_d (1 + B_d - \alpha_i) & \text{if } BF - 1 \ge (F-1) B_d \\ \Upsilon_0 = Z_d \sqrt{\dfrac{M_0 \alpha_0 (1-\alpha_0) B_{\min}}{1 - B_{\min}}} & \text{if } BF - 1 < (F-1) B_d \end{cases} \tag{41}$$

Then, as in the equation (34), Υ follows a Gaussian distribution with the mean of Υ_0 and the variance σ^2. Rewrite the conditions in the equation (34) as,

$$\frac{\Upsilon - \Upsilon_0}{\sigma} \geq c_{i+1} \tag{42}$$

Immediately we see that any SNP participating in the interaction at the worst situation has the probability of $1-\Phi(c_{i+1})$ to satisfy the condition (42). Recall that there are $(d-i)$ ways to grow the (i+1)-order seed from the i-order seed. The probability that no (i+1)-order seed can be generated from the i-order seed is $(\Phi(c_{i+1}))^{d-i}$. So the survival rate for the i-order seed growing is,

$$\zeta_i \geq 1-(\Phi(c_i))^{d-i+1} \tag{43}$$

The "≥" symbol in (43) comes from the fact there are possibly more than one (i-1)-order seed to grow the i-order seeds for a certain d-order interaction.

For the construction of the final interaction candidate as in the formulae (39), we have Υ_0 at the worst situation as the following,

$$\Upsilon_0 = Z_d \sqrt{\frac{M\alpha(1-\alpha)B_1B_2}{1-B_1B_2}}.$$

According to ***Theorem 10***, σ^2 is the variance for Υ in the formulae (39). Again, $(\Upsilon - \Upsilon_0)/\sigma$ follows a standard Gaussian distribution. Thus, we have the survival rate ζ_d satisfying,

$$\zeta_d \geq 1-\Phi(c_d) \tag{44}$$

Similarly, The "≥" symbol is due to the multiple possible ways to construct the final candidates. For example, a three-order interaction [a, b, c] can be constructed in three ways such as [[a,b],c], [[a,c],b] and [[b,c],a]. However, some of them may be eliminated because of the previous seeds growing steps.

Hence, the detection rate at the worst situation is,

$$\begin{aligned}
\tau &= \prod_{i=1}^{d} \zeta_i \\
&\geq \left(1-\Phi\left(c_d\right)\right)\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(c_i\right)\right)^{d-i+1}\right) \\
&= v_d \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(c_i\right)\right)^{d-i+1}\right) \\
&= v_d \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1} c_1\right)\right)^{d-i+1}\right) \\
&= v_d \, f\left(c_1\right) \\
&= v_d \frac{h}{v_d} \\
&= h
\end{aligned} \tag{45}$$

From the property ((30) and (32)) of the worst situation, the number $\Upsilon_0^{'}$ of TGEP associated with any significant interaction is larger than the one at the worst situation, that is, $\Upsilon_0^{'} = \Upsilon_0 + a$, where $a \geq 0$. Then, the survival rate $\zeta_i^{'}$ for any significant interaction can be written as,

$$\zeta_i^{'} \geq 1-\left(\Phi\left(c_i - \frac{a}{\sigma}\right)\right)^{d-i+1} \geq \zeta_i \tag{46}$$

So the detection rate for any significant interaction is larger than τ and hence larger than *h*.

Recalling that there are two possible ways to relax the worst situation as discussed in the section of the definition of the worst situation, one is to increase the frequency of the disease-risk genotype and the other way is to decrease the p-value. When we have a significant interaction

easier to detect than the worst situation, the difference (that is a) between the $\Upsilon_0^{'}$ and Υ_0 at the seeds growing is larger than the difference at the construction of the final interaction candidate, because it relaxes in both ways at the stage of seeds growing (the first stage) and relaxes in the second way only at the stage of the construction of the final interaction candidate (the second stage). Thus, if we set the parameter c_d comparable to be others, the survival rate as in (46) will be much lower for the second stage, which will result in a low overall detection rate. We set a much smaller c_d as in (40) to improve the detection rate. The number 0.95 in (40) is chosen to make sure that we have more than 95% chance to detect a borderline significant interaction at the second stage.

For $d > 2$, there are many possible combinations of $c_i\,(1 \le i < d)$ to satisfy the same detection rate at the worst situation. It is desirable to select the combination of c_i that maximize the detection rate when $a > 0$. However, the best combination depends on the value of a and it is impossible to obtain a universally optimal combination. Setting all $c_i\,(1 \le i < d)$ equal is a choice but obviously not a good choice. Denote σ_i the standard deviation of Υ at the i-order seeds growing. Knowing that the average allele frequency is 0.5, we would expect that $\sigma_i = \sqrt{2}\sigma_{i+1}$. Thus, if we set all c_i ($1 \le i < d$) equal to c, it is possible that $c - a/\sigma_1 >> 0$ and $c - \left(\sqrt{2}\right)^{d-2} a \Big/ \sigma_1 << 0$, which means that $\zeta_1^{'} >> \zeta_{d-1}^{'}$ and it will result in a low detection rate. Although there is no universal optimal solution to maximize the detection rate, we could set $c_i - a/\sigma_i$ having the same sign across different orders and relieve the unbalance among the survival rate at different orders. We set that, for $1 \le i < d$,

$$c_i = \left(\sqrt{2}\right)^{i-1} c_1 \tag{47}$$

Averagely, we will have,

$$c_i - \frac{a}{\sigma_i} = \left(\sqrt{2}\right)^{i-1}\left(c_1 - \frac{a}{\sigma_1}\right) \tag{48}$$

II.1.3.D Miscellaneous techniques to speed the computation

Hierarchical structure to reduce the partial repetitious counting

Because the GWAS studies usually involve a lot of subjects, the counting is the most computational-intensive part in the whole approach. We create a hierarchical structure to reduce the repetitious counting. Suppose S^i is the *i*-order seed and $\{g_v, v = 1,...,V\}$ is the genotypes that possibly participate in an interaction. We set a list R to record the active subject, that is, the subject carrying the genotypes $\{g_v, v = 1,...,V\}$. Here, we utilize the facts that there are only a few genotypes having the potential to participate in an interaction, that is, $V << 3^i$. Then the (*i*+1)-order seed growing only needs to count the subjects in the list R. The list is non-increasing and we expect that the list becomes smaller and smaller with the order increasing. There is an extra computation spent on the creation of the list. However, it is very small compared to the possible reduction on the counting because given a seed the creation of the list computes once and the counting to generate the higher order seed will run around N times (where N is the size of the total SNPs in the study).

Hash table to reduce the redundant search

The SNP subset $\{s_1, s_2, s_3\}$ is considered as the same to $\{s_2, s_1, s_3\}$. It would be a waste of computation to evaluate $\{s_2, s_1, s_3\}$ if $\{s_1, s_2, s_3\}$ has already been evaluated. We use hash table to track the status of any SNP subset. Notice that we do not use the direct table because it is not possible to create such a large table record all possible combinations. For example, it requires 500 Gigabits to restore the two-order combinations for a GWAS study of 1 million SNPs. Actually the direct table would be very sparse if we do have enough memory to create a huge table, since only a small fraction of combinations are really evaluated. The sparse property makes us able to solve the problem by hash table. The hash table only needs the memory comparable to the combinations that are really evaluated. If a SNP subset is already in the hash table, it means this SNP subset is already evaluated and we do not need to evaluate it again. Otherwise, we evaluate this SNP subset and add it to the table. The list technique has the same capacity to restore the evaluated combinations, but it lacks the quick access to the combination's status.

Suppose $\{s_{index\_1}, ..., s_{index\_d}\}$ is a *d*-order SNP subset with the increasing SNP indices. Denoting N the size of the SNPs in the study, we can set a unique number to represent each SNP subset as,

$$\psi = \sum_{i=1}^{d} N^{d-i} Index\_i \tag{49}$$

Assuming the size of the hash table is T, the SNP subset $\{s_{index\_1}, ..., s_{index\_d}\}$ can be restored and accessed at the position,

$$Pos = \left\lfloor T\left(\psi \frac{\sqrt{5}-1}{2} - \left\lfloor \psi \frac{\sqrt{5}-1}{2} \right\rfloor\right)\right\rfloor \tag{50}$$

For more details please refer to the book (Cormen and Cormen, 2001).

Direct table to reduce the duplicated computation

When we calculate the CDF of the hyper-geometric distribution as in the formulae (1), we rely on the computation of the binomial coefficients ("*n* choose *k*"). The binomial coefficients are used repeatedly and a direct table can reduce the duplicated computation. In addition, if we restore all the binomial coefficients in the table at one time, we could further reduce the computation by utilizing the recursive property. The binomial coefficients can be calculated as the following,

$$\binom{n}{k} = \frac{n(n-1)...(n-k+1)}{k(k-1)...1} \tag{51}$$

which would need 2*k* times of multiplication operation. When the binomial coefficients are computed systematically, we could use the following property,

$$\binom{n}{k} = \frac{n-k+1}{k}\binom{n}{k-1} \tag{52}$$

We should point out that, although these techniques are simply adopted from existing literature and there is not too much 'novel' contribution to the techniques themselves, their impacts on the computation reduction are significant. Depending on the specific configuration of the dataset, the saving can be up to several folds.

II.1.3.E Brief summary on heuristic search strategy

The flow chart of the heuristic search is presented in Figure 1.d. The heuristic search is organized around the new concept TGEP - the Transferable Genotype-Effect Potential. One useful characteristic of TGEP is the invariance across different orders, which makes possible the

inference from the lower-order information to the higher-order effect. TGEP also enables the probability framework that links the input requirement of the performance to the observed lower-order statistics.

In the process of devising heuristic search strategy, we made the follow mathematical contributions. (1) We proved that the expectations associated with TGEP are invariant across different order of SNP subset. (2) We showed that there is an algorithm with logarithm complexity to set the parameters based on the performance requirement. (3) We mathematically provided the optimal estimate of the TGEP quantity based the information obtained from the seeds statistics and the marginal statistics. (4) We proved that the given heuristic search procedure controls the lower bound of the performance.

Theorem 10: **Assume g_1 and g_2 are genotypes specified by two non-overlapping SNP subsets. Suppose there are totally M subjects with j subjects being cases. The genotype g_1 specifies a population with r_1 subjects and k_1 cases. The genotype g_2 specifies a population with r_2 subjects and k_2 cases. Write $\alpha = j/M$, $B_1 = k_1/r_1$ and $B_2 = k_2/r_2$. The maximum likelihood estimation of the number of TGEP associated with $g_1 \bigcap g_2$ is $\frac{\sigma_1^2 \Upsilon_1 + \sigma_2^2 \Upsilon_2}{\sigma_1^2 + \sigma_2^2}$ and the variance of the estimation is $\frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$, where $\Upsilon_1 = k_1 - \frac{j-k_1}{M-r_1} r_1$, $\Upsilon_2 = k_2 - \frac{j-k_2}{M-r_2} r_2$, $\sigma_1^2 = \frac{M(1-B_1B_2)\alpha_1(1-\alpha_1)B_1^{'}(1-B_1^{'})}{(1-B_1)^2}$, $\sigma_2^2 = \frac{M(1-B_1B_2)\alpha_2(1-\alpha_2)B_2^{'}(1-B_2^{'})}{(1-B_2)^2}$, $\alpha_1 = \alpha - \frac{\Upsilon_1}{M}$, $\alpha_2 = \alpha - \frac{\Upsilon_2}{M}$, $B_1^{'} = \frac{B_1(1-B_2)}{1-B_1B_2}$, and $B_2^{'} = \frac{B_2(1-B_1)}{1-B_1B_2}$.**

Theorem 11: **Assume $\Phi(\bullet)$ is the CDF of the standard Gaussian distribution. Let $f(x)=\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1}x\right)\right)^{d-i+1}\right)$. Then, $f(x)$ is a strictly decreasing function and $f^{-1}(\bullet)$ can be solved in logarithm complexity by binary search.**

Theorem 12: **Assume P_d and P_{d-1} be the thresholds of experimental-wise significance for the *d*-order and (*d*-1) order interactions, respectively. Suppose S^d is a significant *d*-order interacting SNP subset and g_d is the disease-risk genotype specified by S^d. Denote $F=\left(\Phi^{-1}(P_d)/\Phi^{-1}(P_{d-1})\right)^2$, where $\Phi^{-1}(\bullet)$ is the inverse CDF of standard normal distribution. If the single SNP *s* is included in S^d with *B* being the frequency of disease-risk allele and the (*d*-1)-order subsets of S^d excluding *s* is not significant and, the frequency of the disease risk genotype g_d is larger than $B_{\min}=(BF-1)/(F-1)$.**

II.2 Experimental Results

This subsection shows the results on both simulation and real datasets. We have selected eight peer methods to compare, including three classic methods that do not consider the interaction effects. The eight methods are described in subsubsection II.2.1. Since the accuracy of the type I error is an important measure of the goodness for a statistical method, in subsubsection II.2.2 we provide the empirical evaluation of the type I error for the competing methods and our proposed method. In subsubsection II.2.3 we present extensive experiments on the simulation datasets. The results on real applications are discussed in subsubsection II.2.4.

II.2.1 Existing Methods to Compare

Eight peer methods are briefly described in this subsubsection. Actually the first three methods are routine approaches in current GWAS practice, but they are not interaction-oriented. The purpose of including these three methods is to show that the consideration of interaction effect does matter.

II.2.1.A Pearson's chi-square test

A contingency table as shown in Table 18 can be constructed for each SNP. The quantity χ^2 defined below follows a 2-degree of freedom (*df*) chi-square distribution (Agresti, 2002) under the null hypothesis that SNP and disease status are independent.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\left(n_{ij}-\mu_{ij}\right)^2 \Big/ \mu_{ij} \tag{53}$$

$$\mu_{ij} = n_{i+}n_{+j}\Big/ n_{++} \tag{54}$$

Table 18: Contingency table for Pearson's chi-square Test

	$s = 0$	$s = 1$	$s = 2$	Total
Disease	n_{11}	n_{12}	n_{12}	n_{1+}
Normal	n_{21}	n_{22}	n_{23}	n_{2+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{++}

II.2.1.B Logistic regression

Let $\pi(X)$ denote the probability of getting disease for a subject carrying the genotype *X*. Logistic regression model (McCullagh and Nelder, 1989) considers that the logit function of π is a linear function of genotypes, that is,

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right)=\beta_0+\sum_{i=1}^{d}\beta_i x_i \qquad (55)$$

where $\beta=\{\beta_0,\beta_1,...,\beta_d\}$ are the parameters learned via a (convex optimization) maximum likelihood estimation procedure. Let L_d denote the log likelihood with the estimated $\hat{\beta}$ plugged in. Let L_0 denote the log likelihood corresponding to the estimate of β_0 with the constraint that all the other parameters are set to 0. $G=2(L_d-L_0)$ has asymptotically chi-square distribution with *d* degree of freedom.

II.2.1.C Fisher's exact test

When our method is constrained to single-locus evaluation, it reduces to the Fisher's Exact Test (Agresti, 2002).

II.2.1.D Logistic regression with interaction terms

For each pair (x_i,x_j) in the *d*-locus subset, an interaction term (McCullagh and Nelder, 1989) can be obtained as $\eta_{ij}=x_i x_j$ and there are C_d^2 such interaction terms. Denote $L_{d+interact}$ as the maximum log likelihood when both the marginal *X* loci and the interaction terms are used as predictors. Denote L_d as the maximum log likelihood including only *X* as predictors. $G'=2(L_{d+interact}-L_d)$ asymptotically follows a chi-square distribution with C_d^2 degrees of freedom.

II.2.1.E Full interaction model

There are 3^d genotypes for a subset of *d* SNPs. 3^d dummy variables (Marchini, et al., 2005) are constructed and a logistic regression with 3^d parameters is estimated from the data.

Correspondingly, the decision statistic asymptotically follows a chi-square distribution with $3^d - 1$ degree of freedom.

II.2.1.F Information gain

Consider two SNPs *A* and *B*. Let *C* denote the class label. The information gain of *A*, *B* and *C* is defined as follows,

$$IG(A,B,C) = I(A;B \mid C) - I(A;B) \tag{56}$$

where $I(A;B)$ is the mutual information (Cover and Thomas, 2006) between *A* and *B* and $I(A;B \mid C)$ is the conditional mutual information. A large *IG* gives indication (Moore, et al., 2006) of interaction between A and B.

II.2.1.G Multifactor dimensionality reduction (MDR)

MDR identifies a *d*-locus interaction by a constructive induction algorithm that converts the *d*-dimensional variable into a single variable (Ritchie, et al., 2001). A *d*-locus SNP subset specifies 3^d genotypes or cells. The ratio of the number of cases to the number of controls is estimated within each cell. The cell is labeled as 'high risk' if the ratio exceeds a threshold and otherwise labeled as 'low risk'. In this way, the *d*-dimensional space is reduced to the 1-dimensional space of two statuses – 'high risk' and 'low risk'. The subset is then evaluated using a cross-validation based classification measure (Duda, et al., 2001). Specifically, the data is randomly divided into 10 equal-size folds. The MDR model is constructed based on 9 folds and the remaining fold is used to test the model. The error rate is averaged over all ten folds, each held out in turn.

II.2.1.H Bayesian epistasis association mapping (BEAM)

The SNPs are categorized into three groups. Group 0 contains the SNPs that have no association with the disease. Group 1 contains the SNPs that independently (additively) influence the disease status. Group 2 contains the SNPs that jointly (interactingly) influence the disease status. BEAM generates each SNP's posterior probability of belonging to each of the three groups via Markov chain Monte Carlo (MCMC). A Dirichlet distribution is assumed to be the prior distribution with the parameters set to reflect the prior belief concerning the number of disease-risk markers in the study (Zhang and Liu, 2007).

II.2.1.I SNP harvester (SH)

SNP harvester (SH) (Yang, et al., 2009) is a heuristic search scheme to reduce computational complexity and to detect SNP interactions with weak marginal effects. SH consists of three major steps. In the first step, it removes SNPs with significant main effects. In the second step, it randomly initializes a k-tuple SNP subset and swaps one of the members with one of the remaining SNPs if the statistical score is increased. This step is performed iteratively until convergence. In the third step, an L2-norm penalized logistic regression is applied to post-process the SNP subset obtained in the second step.

II.2.2 Evaluation of Type I Error Rate

The type I error is evaluated on the 1000 replication datasets without any ground-truth SNPs present. Denote N the number of replication datasets, M the number of multiple tests in each dataset, T_i the number of false positives in the *i*th dataset under certain threshold of significance level, $i = 1,...,N$. We have two different definitions for the type I error rate. The first definition,

which follows the idea of family-wise error and measure the chance of there being *any* positives in a dataset, is estimated by,

$$F_A = \frac{\sum_{i=1}^{N} \mathrm{I}(T_i \geq 1)}{N} \ . \qquad (57)$$

where $\mathrm{I}(T_i \geq 1) = 1$ if $T_i \geq 1$ is true and otherwise zero.

The second definition of the type I error rate takes the weight of the number of false positives and can be quantified by,

$$F_B = \frac{\sum_{i=1}^{N} T_i}{N} \ . \qquad (58)$$

Strictly speaking, the first definition is more appropriate. However, some publications, such as BEAM, have adopted the second definition. In addition, the second definition has several nice properties and helps clarify some important points concerning the inaccurate type I error rate of some methods, which will be discussed extensively in the sequel. Hence, we provide results for both two definitions.

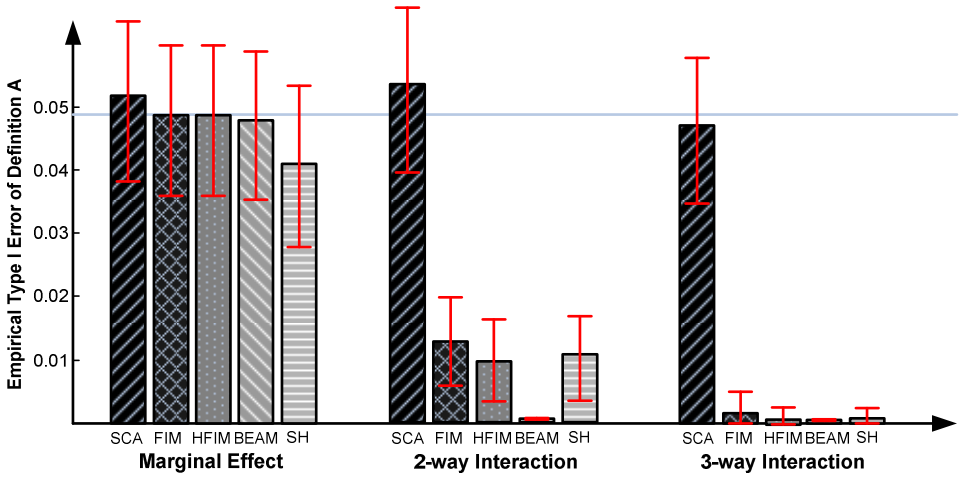

Figure 7: The empirical type I error rates for definition A under different order of effects. The false positives are collected at the 0.05 experimental-wise significance level. The light blue horizontal line indicates the expected type I error rate. The red lines show the 95% confidence interval.

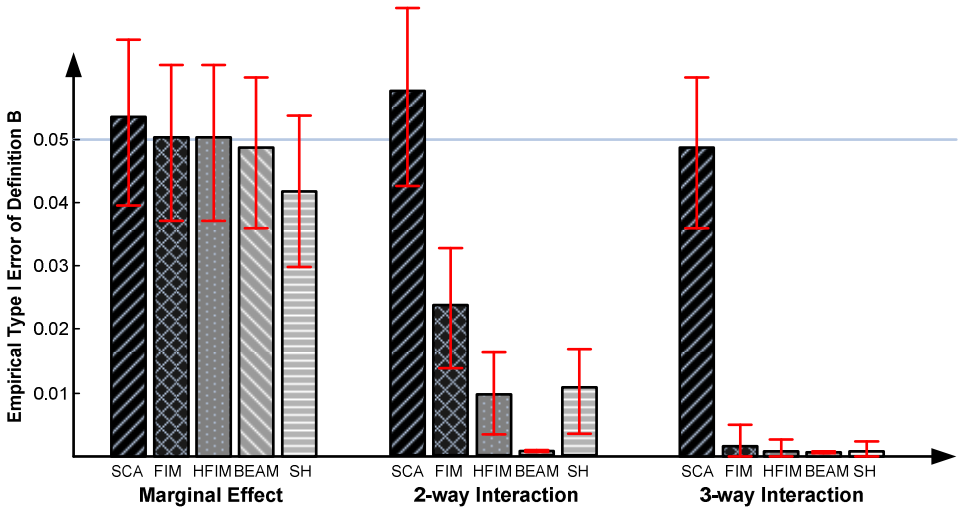

Figure 8: The empirical type I error rates for definition B under different order of effects. The false positives are collected at the 0.05 experimental-wise significance level. The light blue horizontal line indicates the expected type I error rate. The red lines show the 95% confidence interval.

Figure 7 and Figure 8 show the empirical type I error rates under different order of interactions for the first and the second definitions, respectively. The 95% confidence intervals are also labeled in the figures by assuming a binomial model and estimating the variance as $F(1-F)/N$.

Table 19 and Table 20 provide the detailed values in case the reader wants to know the exact empirical type I error rates. Through the figures and the tables, we can clearly see the existing methods such as FIM, HFIM, BEAM and SH are very conservative and the degree of conservativeness increases with the order of interaction. Here, HFIM is a modification to FIM (namely Hierarchical Full Interaction Model) proposed in the BEAM paper. The object of HFIM is to avoid the high false positives FIM incurs disease-associated SNPs are truly present in the data. If some SNPs are detected significant through either marginal effect or interaction effect, HFIM will keep them from participating in higher order of interactions.

Table 19: Type I error for definition A and its 95% confidence interval. The false positives are collected at the 0.05 experimental-wise significance level.

	FIM	HFIM	BEAM	SNPHarvester	SCA
Marginal effect	0.049 [0.0356, 0.0624]	0.049 [0.0356, 0.0624]	0.048 [0.0348, 0.0612]	0.041 [0.028, 0.0533]	0.052 [0.0382, 0.0658]
2-way interaction	0.013 [0.006, 0.0200]	0.010 [0.0038, 0.0162]	0.00 [0.00, 0.00]	0.011 [0.0045, 0.0175]	0.054 [0.0400, 0.0680]
3-way interaction	0.002 [0.0, 0.0048]	0.001 [0.0, 0.003]	0.00 [0.00, 0.00]	0.001 [0.0, 0.003]	0.047 [0.0339, 0.0601]

Table 20: Type I error for definition B and its 95% confidence interval. The false positives are collected at the 0.05 experimental-wise significance level.

	FIM	HFIM	BEAM	SNPHarvester	SCA
Marginal effect	0.050 [0.0365, 0.0635]	0.050 [0.0356, 0.0624]	0.049 [0.0356, 0.0624]	0.042 [0.0296, 0.0544]	0.054 [0.0400, 0.0680]
2-way interaction	0.024 [0.0145, 0.0335]	0.010 [0.0038, 0.0162]	0.00 [0.00, 0.00]	0.011 [0.0045, 0.0175]	0.058 [0.0435, 0.0725]
3-way interaction	0.002 [0.0, 0.0048]	0.001 [0.0, 0.003]	0.00 [0.00, 0.00]	0.001 [0.0, 0.003]	0.049 [0.0356, 0.0624]

II.2.2.A Sources of conservativeness of the peer methods

There are several sources that cause the conservativeness, some which have already been explored in the comparison paper (Chen, et al., 2011). We summarize the following four factors for the conservativeness: small sample size and discreteness, heuristic search, correlation of multiple tests and multiple orders of interaction.

a) small sample size and discreteness

Although the total sample size is usually large in the GWAS studies, some cells in the contingency table may have small sample size. The small sample size combined with the discrete property of the data makes the significance assessment of the summary statistics conservative in the existing methods, especially when we are interested in small p-values. Two phenomena will occur in the case of discrete data and limited sample size: (1), **non-infinitesimal p-value** (that is, infinitesimal p-value does not exist with the distribution), and (2) **discrete p-value**. These two phenomena are genuine and make the significance assessment inaccurate regardless of whether the sample size is large or small; however, the small sample size will exaggerate the inaccuracy. We have examined this accuracy in the comparison paper (Chen, et al., 2011). The results clearly show the conservativeness of the summary statistics in the existing methods. In addition, we have observed that the conservativeness increases with smaller p-value, which is consistent with our explanations as the impact of the non-infinitesimal p-value and discrete p-value becomes more severe with smaller p-value.

b) heuristic search

Let p_{ij} be the p-value of the j^{th} test in the i^{th} replication. We further use s_{ij} to indicate whether the j^{th} test in the i^{th} replication is searched and evaluated. $s_{ij}=1$ denotes the searched status and $s_{ij}=0$ denotes the unsearched status. Then, the number T_i of false positives in i^{th} replication is determined by both the search status and the p-value for each test. With the Bonferroni corrected significance level α, we have $T_i=\sum_{j=1}^{M} s_{ij}\mathrm{I}\left(p_{ij}\le\frac{\alpha}{M}\right)$ and can rewrite the second definition of the type I error rate as,

$$F_B=\frac{\sum_{i=1}^{N}T_i}{N}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}s_{ij}\mathrm{I}\left(p_{ij}\le\frac{\alpha}{M}\right)}{N} \qquad (59)$$

We have the following theorem about the expectation of F_B.

Theorem 13: **Denote X_{ij} the j^{th} random variable in the i^{th} replication with the p-values p_{ij} associated with each random variable. Assume X_{ij} has identically distribution. Writing $\Pr(s=1|\,p\le\alpha/M)$, we have $\mathbf{E}(F_B)=\alpha\times\Pr(s=1|\,p\le\alpha/M)$.**

$\Pr(s=1|\,p\le\alpha/M)$ represents the probability that a test is searched and evaluated if the test is significant after Bonferroni correction. If a heuristic search is not effectively designed to search the significant tests, we could expect $\Pr(s=1|\,p\le\alpha/M)$ to be small and thus the estimated type I error rate will be very conservative compared to the claimed significance level.

c) correlation of multiple tests

The correlation among multiple tests has been **long and widely** blamed for the conservativeness observed in the type I error rate. For example, the BEAM noticed the

conservativeness with their methods. After removing the effect of heuristic search (they implemented an exhaustive search), the conservativeness is still there. Finally, they concluded that this was due to the correlation among the couplets caused by the shared SNPs in some pairs of couplets. However, with careful examination, we have found that **correlation does not contribute to the conservativeness at all**, as shown by the following corollary.

Corollary 13.1: **The correlation in the multiple tests does not contribute to the conservativeness in the sense that** $\mathbf{E}(F_B) = \alpha$ **if the tests are exhaustively searched.**

This conclusion is obtained under the second definition of the type I error rate. Does the correlation impact the conservativeness under the first definition? The answer is YES. Fortunately, there is a simple approach to assess the degree of impact, as illustrated by the following theorem:

Theorem 14: **For a large number of independent multiple tests,** $\mathbf{E}(F_A) \approx \frac{1-\mathbf{e}^{-c\alpha}}{c\alpha}\mathbf{E}(F_B)$**,** **where** α **is the Bonferroni corrected significance level and** $c = \Pr(s=1 \mid p \le \alpha/M)$**. Especially, when** $\alpha = 0.05$ **and** $c = 1$**,** $\mathbf{E}(F_A) \approx 0.98\,\mathbf{E}(F_B)$**.**

Thus, the closeness between F_A and F_B can be considered as a criterion to measure whether the multiple tests behave independently. From this criterion, we can see that the

multiple tests in the methods including ours all look like independent tests (except for FIM), based on the observed closeness between F_A and F_B in Table 19 and Table 20.

d) cross-order comparison

In the practice of SNP interaction analysis, multiple orders of interaction are simultaneously examined. To promise that the discovered SNP group has true interaction effect and to avoid the occurrence that the significance of a d-SNP group is caused by its subgroup, a seemingly natural and widely accepted scheme is to put constraints on its subgroup's significance, that is, the p-value of the subgroup of an interacting d-SNP group should be larger than a certain value. This very operational procedure turns out to have very significant theoretical impact on the accuracy of the significance evaluation of the d-SNP group. Putting it straightly, this scheme makes the p-value evaluation **systematically biased to be conservative**. We would like to emphasize that although this approach has been practiced a long time and taken for granted as being a reliable procedure, the fact that this introduces systematic bias **has never been pointed out and explored until now**.

Denote P_G the p-value (unconditional) associated with the d-SNP group. Similarly, P_S is used to denote the p-value associated with its sub-group. Suppose τ is the threshold that is required for the sub-group's significance to surpass. If the observed p-value for the d-SNP group is v, the actual p-value for such an event given the constraint under consideration should be calculated as $\Pr(P_G \le v | P_S \ge \tau)$, instead of $\Pr(P_G \le v)$.

From the following ***Theorem 15***, we can clearly see the conservativeness due to ignoring the constraint on the sub-group, in the process of evaluating the actual p-value, as the positive dependence (Tong, 1990) can be easily established by the fact that a higher significance of the d-SNP group usually implies a higher significance of its subgroup.

Theorem 15: **If** P_G **and** P_S **are two positively dependent variables,** $\Pr(P_G \le v | P_S \ge \tau) \le \Pr(P_G \le v)$**.**

II.2.2.B Our solution to avoid the conservativeness within the SCA framework

All those factors discussed above have been taken into account in our SCA framework. Correspondingly, our algorithm improves on the following aspects,

(1) A hypergeometric distribution is applied, which is an exact distribution regardless of the sample size.

(2) A special trick is designed to cope with the non-infinitesimal p-value, by only considering the cells with small-enough p-values.

(3) An algorithm called mid-p-value is adopted to relieve the effect of discrete p-value.

(4) A theoretically sound heuristic search strategy is created to touch most of the possible significant candidates.

(5) The actual p-value with the constraint of subgroup significance is explicitly computed.

II.2.3 Results on Simulation Data

The generation of the simulation data is discussed in subsubsection II.2.3.A. Totally seven genetic models are inserted in the simulation data. The subsubsection II.2.3.B shows the testing results on detection power of our proposed method compared to the peer methods. The efficiency of the heuristic search is investigated in subsubsection II.2.3.C. The subsubsection II.2.3.D presents experiments for identifying interaction effects among marginally significant SNPs and comparison to logistic regression with interaction terms.

II.2.3.A Description of the simulation data

The simulation data, corresponding to approximately 1000 cases and 1000 controls with 1000 SNPs, are generated from the real data on 223 individuals that were genotyped on the 317K Illumina HumanHap300 BeadChip as part of the New York City Cancer Control Project (NYCCCP). Seven models as the possible causes of the disease are inserted and 17 ground-truth SNPs are associated with these seven models.

The simulation of these data proceeds as follows. Consider a matrix with 223 rows corresponding to NYCCCP individuals and 317,503 columns corresponding to the 317,503 SNPs. The elements of this matrix are the individual genotypes. Partition the columns into “bins” of 500 consecutive SNPs. That is, 636 bins, where the last bin only has 3 SNPs. The simulated genome scan data for each individual is obtained by random draws (with replacement) from real data matrix of 223 individuals and 636 bins of 500 SNPs. Specifically, the simulated data for an individual is generated by randomly selecting the first bin (first column) from the 223 individuals (rows), randomly selecting with replacement the second bin from the 223 individuals, randomly

selecting with replacement the third bin from the 223 individuals, and so on through all 636. Thus the data retains the basic patterns of linkage disequilibrium (Pasternak, 2005), missing data, and allele frequencies as that observed in the original genome scan data. The exception to this is only at the 635 breaks in the genome corresponding to the bin boundaries. The same random process of sampling bins with replacement is repeated for all individuals in the resulting simulated dataset.

The data was simulated under the alternative hypothesis described below and with no missing data. Seven models were explored in the simulation. Five of the seven models involve interactions. None of the interactions are motivated by an additive or multiplicative model. Rather, all interactions are generated by more complex Mendelian inheritance patterns. A total of 17 SNPs influence disease status. For each of the analytic models the total number and which individual SNPs were selected to be retained (which SNPs were detected) were recorded.

Model 1 – Five locus interaction among common alleles, nearly fully penetrant.

The model assumes a five-locus interaction under a dominant genetic model for the minor (less frequent) allele at each locus. It assumes a minor allele frequency of 0.30 at each locus. The expected number of such genotype combinations is approximately 35 per 1000 subjects under complete independence. The penetrance function (probability of disease) is zero if the minor allele is not present at each locus and 0.90 if the minor allele is present at each locus. In equation form it is:

$$\text{Prob}\left(\text{Disease}\left|G^{A}_{12\,or\,22} \wedge G^{B}_{12\,or\,22} \wedge G^{C}_{12\,or\,22} \wedge G^{D}_{12\,or\,22} \wedge G^{E}_{12\,or\,22}\right.\right) = 0.90, \text{ zero otherwise.} \qquad (60)$$

Here, G_{12} indicates the genotype is 12 (i.e., heterozygous). The five loci are denoted by the superscript, and ^ denotes intersection. Note: $G_{12}=G_{21}$. The indexes of these five SNPs in the simulation data are 340, 877, 931, 962 and 994.

Model 2 – Three locus interaction among reasonably common alleles, fully penetrant

The second model assumes a three-locus interaction. The minor allele frequencies at the three loci are 0.25 for A, 0.20 for B, and 0.20 for C. The model is fully penetrant (i.e., probability of disease is one given the predisposing genetic factors). In table format we have:

		G^A_{11}				G^A_{12}				G^A_{22}		
		G^C				G^C				G^C		
		11	12	22		11	12	22		11	12	22
	11	0	0	0		0	0	0		0	1	1
G^B	12	0	0	0		0	0	0		1	1	1
	22	0	0	0		0	0	0		1	1	1

Here, the set of three columns under G^A_{11} form a sub-table that corresponds to the possible genotypes at the other two loci G^B and G^C. In the first two sub-tables, the penetrance function or probability of disease is zero. Thus, only in the last sub-table corresponding to the recessive model of locus A is there a nonzero penetrance function. Specifically, it is a fully penetrant model under a recessive model for locus A and dominant models for loci B and C. The indexes of these three SNPs in the simulation data are 917, 891 and 999.

Model 3 – Three locus interaction, common alleles, incomplete penetrance

The third model assumes a three-locus interaction. The minor allele frequencies at the three loci are 0.40 for A, 0.25 for B, and 0.25 for C. The indexes of these three SNPs in the simulation data are 992, 233 and 630. Model 3's penetrance functions can be summarized as:

Prob(disease | heterozygous for at least two loci and not G^A_{11})=0.5.

Prob(disease | homozygous for minor allele at two loci and not G^{A}_{11})=1.0.

		G^{A}_{11}				G^{A}_{12}				G^{A}_{22}		
		G^{C}				G^{C}				G^{C}		
		11	12	22		11	12	22		11	12	22
	11	0	0	0		0	0	0		0	0	0
G^{B}	12	0	0	0		0	0.5	0.5		0	0.5	1
	22	0	0	0		0	0.5	1		0	1	1

Model 4 – Two locus interaction, common alleles, incomplete penetrance, dominant model

The fourth model assumes a two locus interaction for common alleles under a dominant genetic model at each locus. The minor allele frequencies are 0.20 for locus A and 0.30 for locus B. The indexes for A and B are 729 and 972. The penetrance function is summarized in the following table.

		G^{B}		
		11	12	22
	11	0	0	0
G^{A}	12	0	0.75	0.75
	22	0	0.75	1.0

Model 5 – Two locus interaction under a dominant model for the major allele.

The fifth model assumes a two locus interaction under a dominant model for the major allele. The model is for a very common but low penetrant allele. The minor allele frequencies at these two loci are both 0.25. The indexes for A and B are 594 and 852. The penetrance function is summarized in the following table.

		G^{B}		
		11	12	22
	11	0.10	0.10	0
G^{A}	12	0.10	0.10	0
	22	0	0	0

Model 6 – Single locus dominant model, partial penetrance of an uncommon allele.

Model 6 assumes a partially penetrant dominant model at one locus with a minor allele of 0.10. The penetrance function is Prob(disease | G_{12} or G_{22}) = 0.5 and zero otherwise. The index for G is 993.

Model 7 – A single locus model under a recessive genetic model and a common allele.

Model 7 assumes a partially penetrant recessive model at one locus with a minor allele of 0.40. The penetrance function is Prob(disease | G_{22}) = 0.5 and zero otherwise. The index for G is 865.

Each model acts independently of the others and generates a phenotype label. For a certain subject, if *any* model (randomly) generates positive disease status, the subject is finally labeled as a disease subject.

II.2.3.B Detection power after incorporation of interaction effects

The performance of SCA-HCIG was tested on simulation datasets and compared to eight reference methods. The experimental results show very promising relative performance of SCA-HCI, with considerably high sensitivity and specificity, quantified by both the sensitivity at allowable false positive rate (FPR) and the area A_z under the receiver operating characteristic (ROC) curve.

In our comparative studies, we considered eight widely-used and most relevant existing methods, namely, Pearson Chi-square Testing (Chi2) (Stacey, et al., 2007), Logistic Regression (LR) (Scott, et al., 2007), Fisher's exact test (FET) (Duerr, et al., 2006), Logistic regression with

interaction terms (LRIT) (Harley, et al., 2008),Full interaction model (FIM) (Marchini, et al., 2005), Information gain (IG) (Moore, et al., 2006), Multifactor Dimensionality Reduction (MDR) (Ritchie, et al., 2001), and Bayesian Epistasis Association Mapping (BEAM) (Zhang and Liu, 2007) as described in the previous subsection. We re-implemented the algorithms for the first six methods and adopted the authors' public codes for the last two methods, in order to compare the relative performance of the competing methods.

The first three methods (Chi2, LR and FET) evaluate the SNPs individually with no parameter setting required. The next three methods (LRIT, FIM and IG) encode the interaction effects up to the second order. The search schemes for these three methods are not considered here -- we used the exhaustive search to fully explore the sensitivity and the specificity of these methods. The last two methods (MDR and BEAM) are much more sophisticated and complicated approaches which leave several parameters to the users to determine. The latest version for MDR (2.0 Beta 6) is adopted in our experiments. The attribute range is set from 1 to 5, that is, interactions are searched up to the fifth order. A 10-fold cross validation is selected. The random search type is chosen and the running time was set to be 10 hours. We used the default filter selection 'ReliefF' and all the parameters associated with the filter were set to their default values. We downloaded the BEAM software *dos* version from the authors' website with the package created on Sep.18th 2007. There are 9 parameters in total for BEAM. Except for the number of MCMC chains, all the other parameters are set according to the suggested values by the authors. We set the number of MCMC chains to be 100 instead of the suggested number 1, because we observed the randomness for different runs if the number of chains is small. There are two parameters for our SCA-HCIG: the maximum order of interactions and the detection rate at the worst situation. In

the experiments, we searched up to the fifth-order interactions and the detection rate at the worst situation was set to be 0.1.

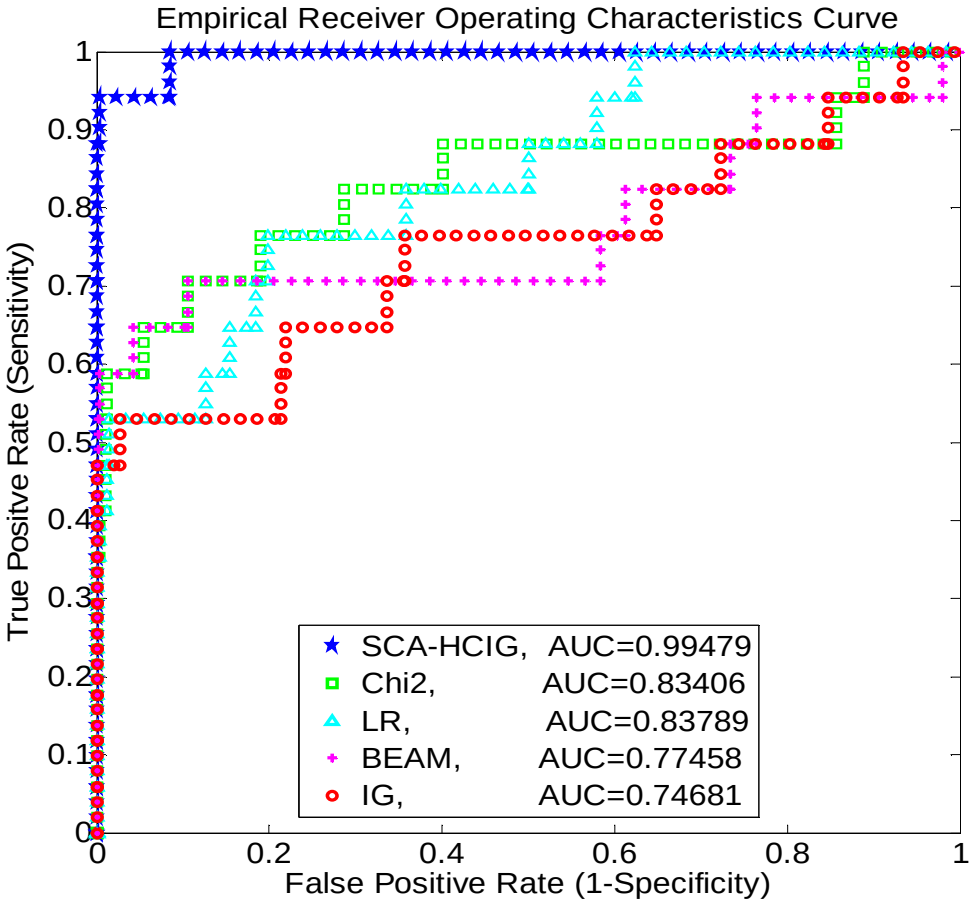

Figure 9: ROC curves of relative detection performance tested on 1000-SNP simulation.

We tested all nine methods including SCA-HCIG on the same simulation data, and subsequently calculated the empirical ROC curves based on the known ground truth and using varying detection thresholds. Figure 9 shows the empirical ROC curves of the relative performance by five methods and the corresponding A_z, tested on the 1000-SNP simulation data. It can be clearly seen that SCA-HCIG significantly outperforms existing methods in that SCA-HCIG has the area A_z more than 0.99, while all the other existing methods have A_z less than 0.84.

Table 21: Relative performance of 9 competing detection methods.

	SNP Index	Chi2	LR	FET	LRIT	TFI	IG	MDR	BEAM	SCA-HCIG
M1	340									√
	877									√
	931									√
	962									√
	994									√
M2	917	√	√	√	√	√	√		√	√
	891						√			√
	999						√			√
M3	992									√
	233	√	√	√	√	√	√		√	√
	630									√
M4	729	√	√	√	√	√	√	√	√	√
	972	√	√	√	√	√	√		√	√
M5	594									
	852									
M6	993	√	√	√	√	√	√	√	√	√
M7	865	√	√	√	√	√	√		√	√
Detection rate with no false positives		35%	35%	35%	35%	35%	47%	12%	35%	**88%**
Execution time (hours, minutes, seconds)		<1s	1.7s	<1s	49.8s	4.5m	2.3m	10h	2.8h	49.7m

Note that in marker discovery with candidate marker set as large as in GWAS studies, the extreme left part of the ROC is the most important and informative, especially the segment that that has no false positives. Table 21 shows the detection rate with zero false positive and the detailed detection on each model. Chi2, LR, FET, LRIT, BEAM and SCA-HCIG all have p-value outputs and their detection rates are calculated based on the detected markers which have Bonferroni corrected p-values less than 0.05. Although FIM also claims to output p-values, we have observed that its Bonferroni corrected p-values are not reliable and the threshold of 0.05 results in a lot of false positives. Instead, we use the original p-values to rank the markers and the detection rate is obtained by retaining the largest set of top markers without false positives. The

IG method outputs its information gain score and the original author did not provide the corresponding p-value. Similar to FIM, we use the information gain score to rank the markers and compute the detection rate based on the largest set of top markers without false positives. MDR selects the best model with the smallest testing error. This contains two ground-truth SNPs and one false positive in our experiment.

As in Table 21, SCA-HCIG achieves a marker detection rate of 88%, compared to the best others of 47%, the best rate among the other methods. M1 is a fifth-order interaction model and shows very little marginal effects. It is not surprising that Chi2, LR and FET cannot detect it since they are originally designed to catch marginal effects. None of the compared methods except for ours can detect any SNP in the fifth-order model. This model indicates SCA-HCIG has the capability to mine high order interactions with very small marginal effects. M2 is a three-way interaction model and one of the three members shows strong marginal effects. Both SCA-HCIG and IG can detect all three members. But all the other methods only detect the member with strong marginal effects. M3 is also a three-way interaction model, but with a different penetrance table. Like M2, one of the three members in M3 shows somewhat high marginal effect. SCA-HCIG is the only method that can detect the other two members beyond the one with large marginal effect. Both M2 and M3 show our method is also able to detect the interaction models with partial marginal effects. M4 is a two-way interaction model that has strong marginal effects for both its two members. Almost all the methods successfully detect them. However, only SCA-HCIG and BEAM point out these two SNPs interact with each other. M5 is a two-way interaction model with very weak penetrance and large frequency of the risk-allele. The sample size is not large enough to render the power to detect it, so none of the methods discovers it. M6 and M7 are two

individual SNP models with very strong marginal effects and almost all the methods find them. The six SNPs which show strong marginal effects can be detected by almost every method. MDR is an exception and detects only two SNPs. The unsatisfactory performance of MDR may partially come from the strategy that it only carries the best classifier in the training phase to the testing phase. A complex disease may have multiple disease causes (heterogeneity) and thus permits multiple good predictors. Selecting only the best predictor to be tested would lose power to detect other good predictors. IG obtains the best detection rate among the methods, excluding SCA-HCIG. However, IG only provides a score to rank the SNPs, which will be very hard to use in practice because the user simply does not know how to set the threshold to claim significant findings.

To assess whether the superior performance of SCA-HCIG is purely by chance, we generated ten more replicated simulation datasets. Figure 10 gives the box plot of the detection results on the ten datasets. The detailed detection results for each dataset are presented in Table 22. The detection rate is calculated in the same ways as in Table 21. The red line at the center of the box is the average detection rate and the box illustrates the range within 1 standard deviation. Clearly, SCA-HCIG has significantly better detection rate than all the other methods and is consistently better than the others on all datasets. One point deserving to note here is that IG is not significantly better than other methods based on the results in Table 22. Actually, no one of the existing methods is significantly better than others, since the boxes overlap with each other. Figure 11 shows the power for each model and for all the ground-truth SNPs when the positives are determined at the 0.05 Bonferroni corrected significance level. Figures 12-17 show marker detection power evaluation for varying false positives, for each individual model and for the

whole ground-truth SNP set. From these figures, we can see that SCA outperforms the 5 peer methods on all interaction models except for the fifth model on which the detection power is too low to differentiate any method.

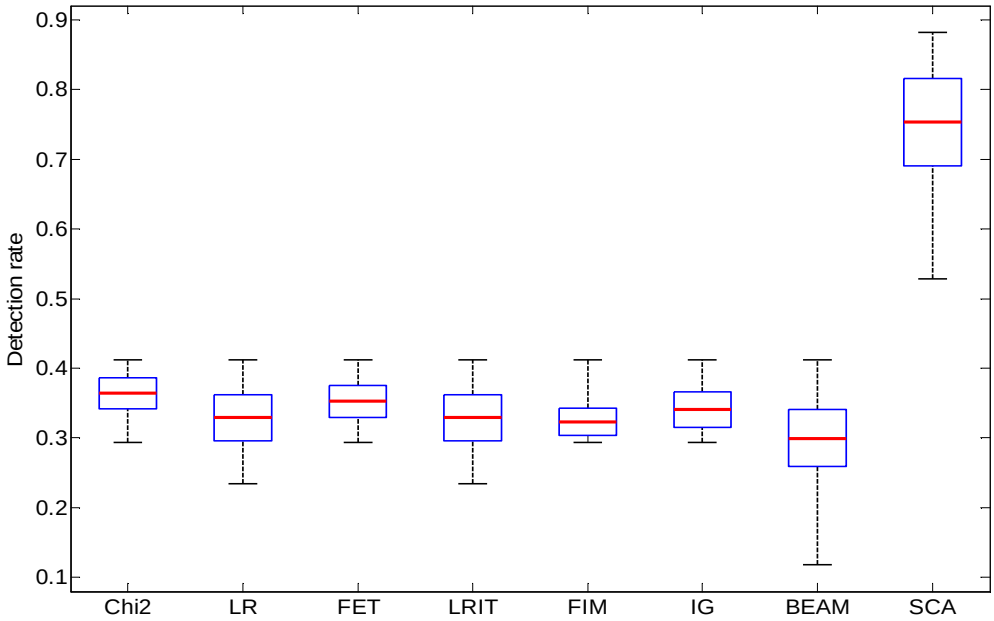

Figure 10: Box plot of detection rate with zero false positives on the 1000-SNP data.

Table 22: Details of detection rate with zero false positives on the 1000-SNP data.

	Chi2	LR	FET	LRIT	FIM	IG	BEAM	SCA
Set1	0.352	0.294	0.294	0.294	0.294	0.294	0.294	**0.765**
Set2	0.294	0.235	0.294	0.235	0.294	0.294	0.235	**0.882**
Set3	0.352	0.294	0.352	0.294	0.294	0.294	0.294	**0.765**
Set4	0.352	0.235	0.352	0.235	0.352	0.294	0.294	**0.882**
Set5	0.412	0.352	0.412	0.352	0.352	0.412	0.412	**0.706**
Set6	0.352	0.412	0.352	0.412	0.294	0.412	0.294	**0.588**
Set7	0.412	0.412	0.412	0.412	0.352	0.412	0.352	**0.529**
Set8	0.412	0.352	0.352	0.352	0.412	0.352	0.294	**0.882**
Set9	0.294	0.294	0.294	0.294	0.294	0.352	0.118	**0.882**
Set10	0.412	0.412	0.412	0.412	0.294	0.294	0.412	**0.647**
mean	0.364	0.329	0.353	0.329	0.323	0.341	0.300	**0.753**
std	0.047	0.069	0.048	0.069	0.042	0.054	0.085	0.132

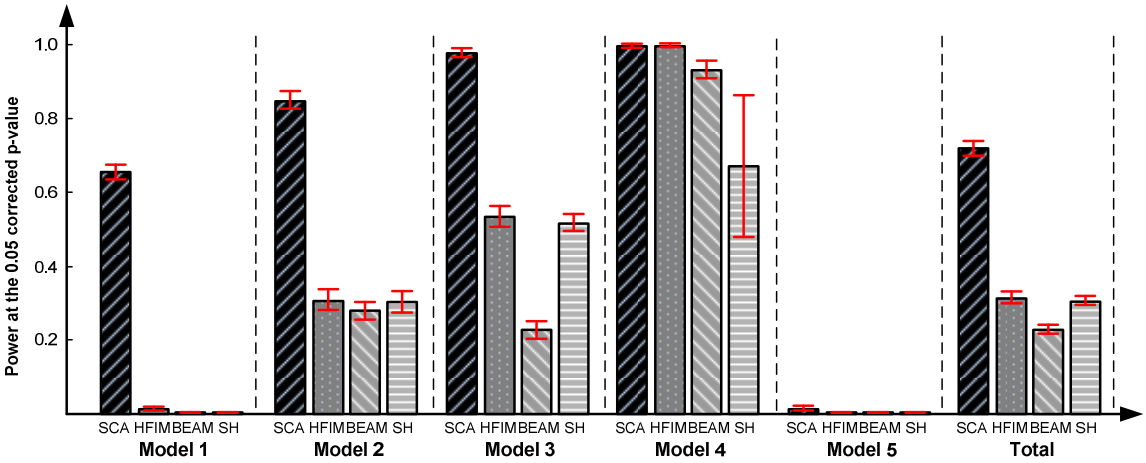

Figure 11: Power at the 0.05 Bonferroni corrected significance level.

Table 23: Power at 0.05 experimental-wise significance level and its 95% confidence interval

	HFIM	BEAM	SNPHarvester	SCA
Model 1	0.022 [0.0154, 0.0286]	0.002 [0.00, 0.004]	0.0040 [0.0012, 0.0068]	0.654 [0.6327, 0.6753]
Model 2	0.3367 [0.3094, 0.3640]	0.2833 [0.2573, 0.3093]	0.3333 [0.3061, 0.3605]	0.8533 [0.8329, 0.8738]
Model 3	0.5367 [5079, 0.5655]	0.2300 [0.2057, 0.2543]	0.5200 [0.4912, 0.5488]	0.9767 [0.9680, 0.9854]
Model 4	1.00 [1.00 1.00]	0.935 [0.9176, 0.9524]	1.00 [1.00 1.00]	1.00 [1.00 1.00]
Model 5	0.0 [0.0, 0.0]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.01 [0.0030, 0.0170]
Total	0.3153 [0.3033, 0.3273]	0.2280 [0.2172, 0.2388]	0.3053 [0.2934, 0.3172]	0.7187 [0.7071, 0.7303]

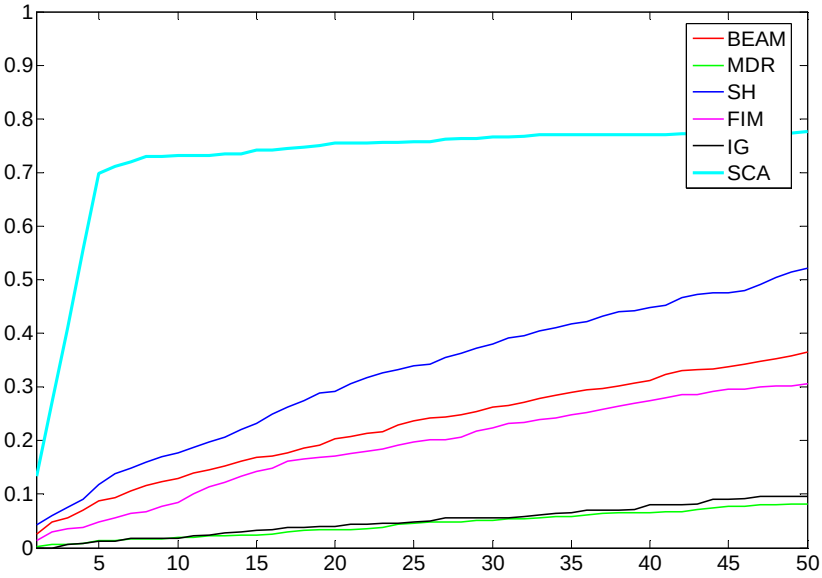

Figure 12: Power versus top selected SNPs on Model 1

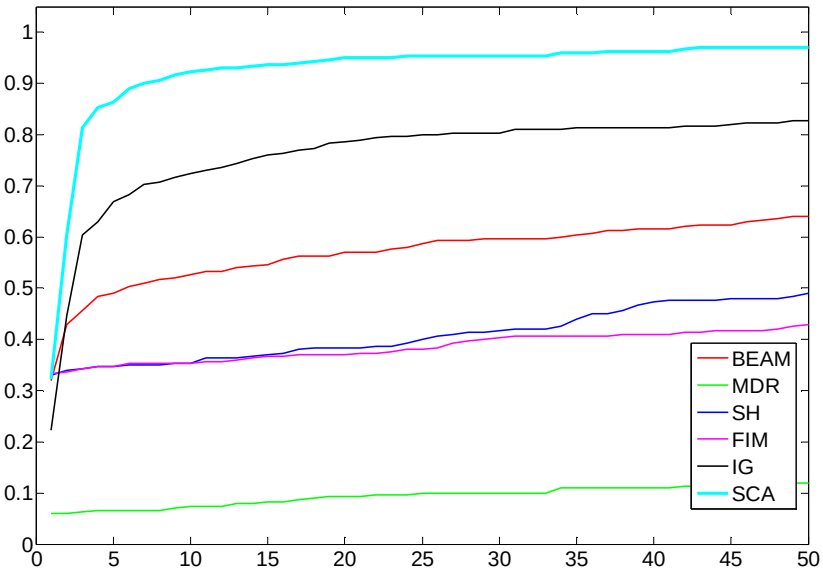

Figure 13: Power versus top selected SNPs on Model 2

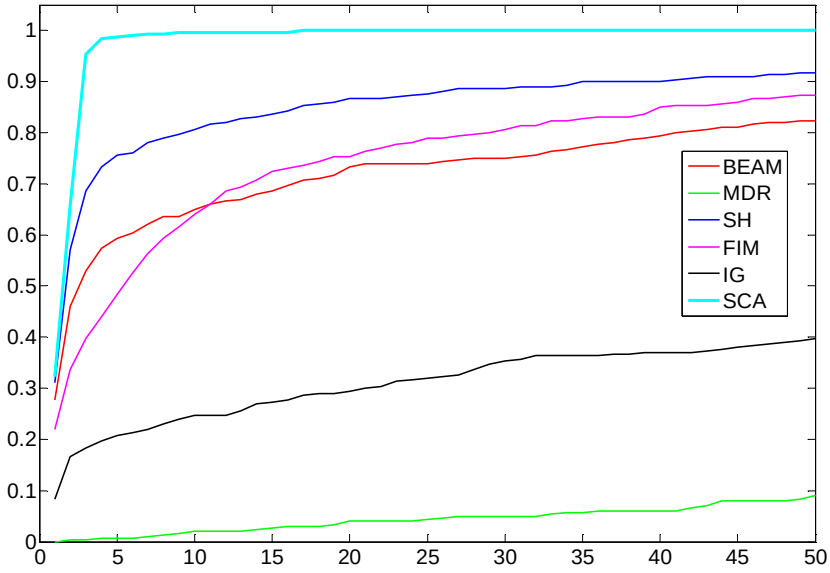

Figure 14: Power versus top selected SNPs on Model 3

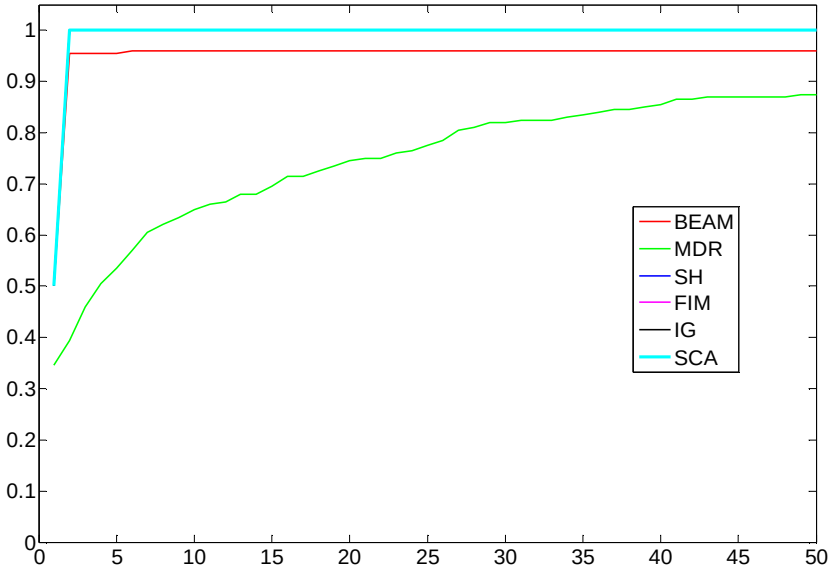

Figure 15: Power versus top selected SNPs on Model 4

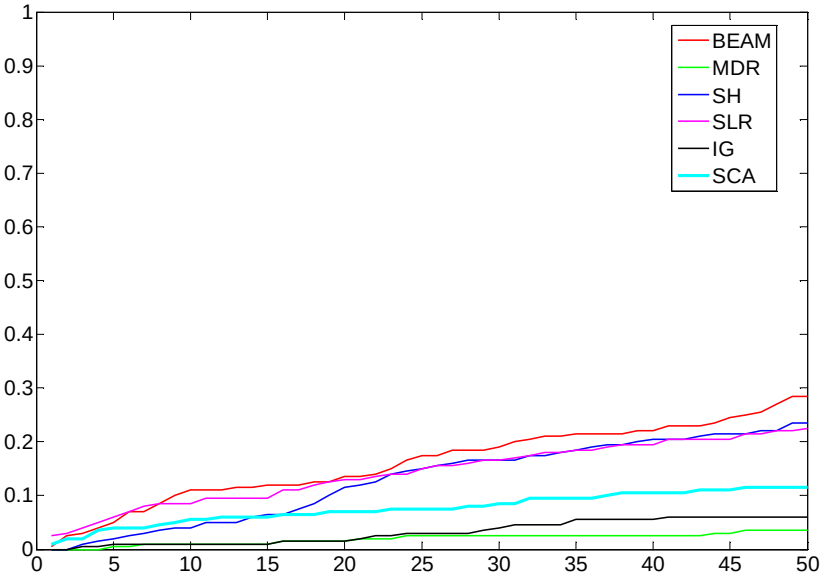

Figure 16: Power versus top selected SNPs on Model 5

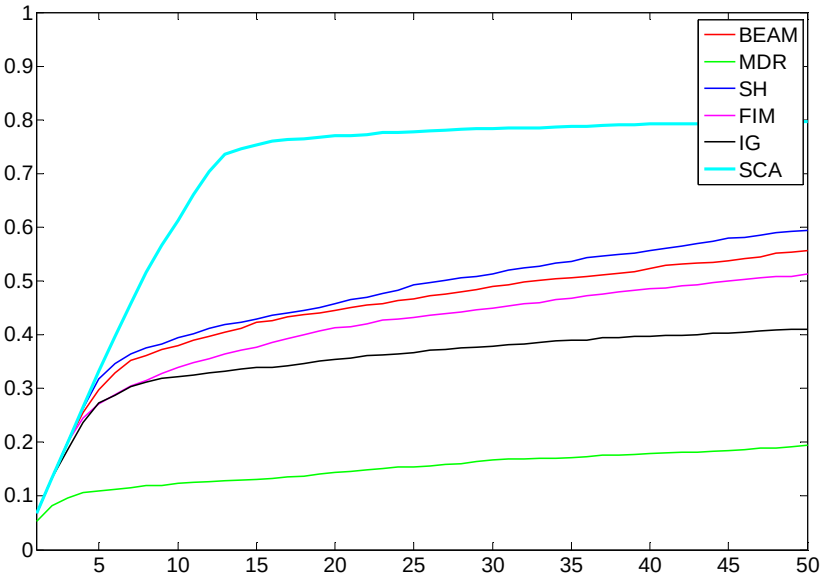

Figure 17: Power versus top selected SNPs on the overall model

II.2.3.C Efficiency of heuristic search

The execution time for each method is also shown in Table 21. As expected, the single SNP based methods such as Chi2, LR and FET are the most computationally-efficient, of which LR takes a little bit more time than the other two since LR involves a complicated learning algorithm. LRIT, FIM and IG are at the scale of minutes. MDR, BEAM and SCA-HCIG are at the scale of hours because these three methods have the capability to discover high-order interactions. While SCA-HCIG has the highest detection rate, it is not the most computationally-expensive. The lesser computation of SCA-HCIG than MDR and BEAM suggests that its superior results are not only due to its elaborate design of heuristic search but also because of the reasonable criterion it uses for measuring the interacting SNP subset.

Figure 18 demonstrates the detection rates and the running times with respect to different values of the heuristic search parameters. The curves of the detection rate and the running time are both non-decreasing functions with the independent variable, as more combinations are searched with a larger low-bound of the performance, and a genuinely significant SNP subset has a better chance to be discovered. Recalling that the worst case rarely occurs and a real marker in practice is much better than the worst situation, the detection rate jumps at a very small value of the heuristic search parameter, which also suggests that a small value is good enough to get a good detection rate. In our program we have set 0.1 as the default value.

The running time increases very rapidly with the heuristic search parameter. Especially when the parameter approaches to 1, the running time is more than exponentially increasing. This is actually a good thing because, based on Figure 18, it means a big deduction of running time can

be obtained with very little loss in performance. Table 24 presents the detailed detection rates and running times with different parameter values. The experiments are implemented and the running time is recorded when the parameter is less than or equal to 0.3. The running time is estimated based on a 100-SNP simulation when the parameter is larger than 0.3. If we set the parameter to be 1.0, which is equal to exhaustive search, the 1000-SNP simulation data is estimated to take about 363,930 years to search up to fifth-order interactions. When we set the parameter to be 0.1, we only need about 50 minutes to finish the analysis. This is a huge improvement of more than 3 billion-fold saving. More interestingly, with the parameter set to 0.95, which means there is at most 5% chance to miss the target, the computation can be accelerated by more than ten thousand fold.

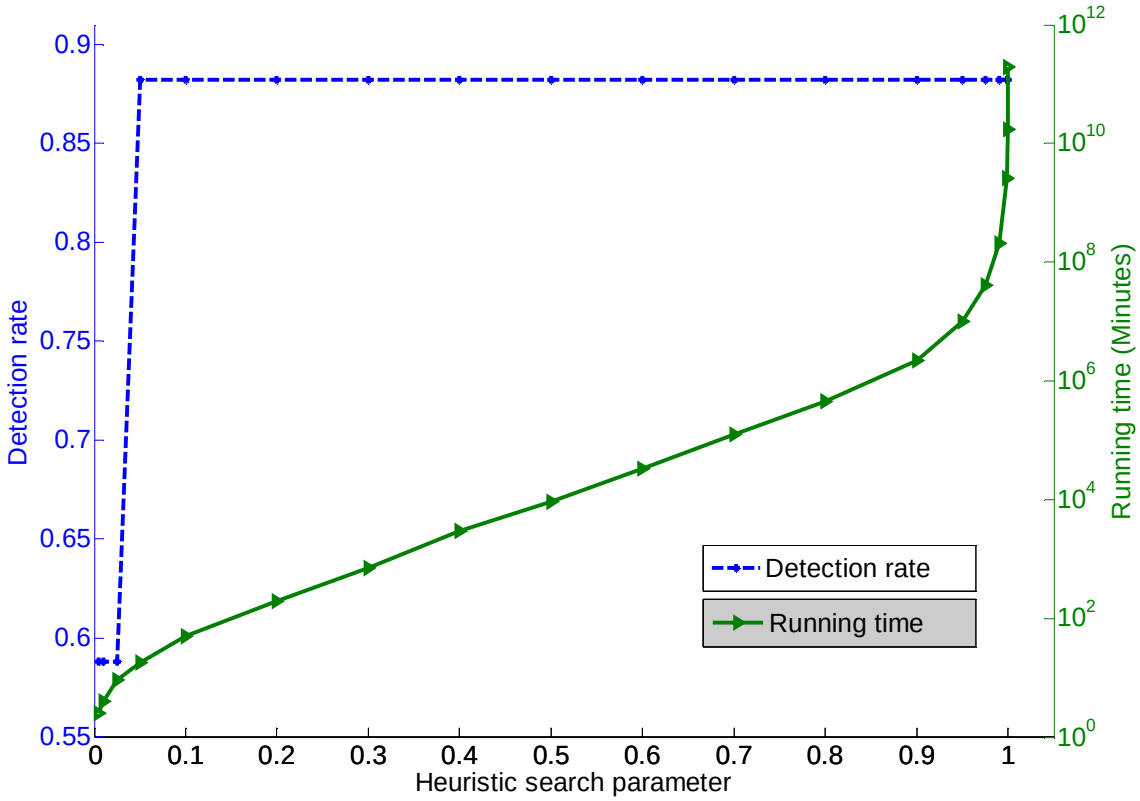

Figure 18: The detection rate and running time with different heuristic search parameters.

Table 24: Details of detection rate and running time with different heuristic search parameters, (m=minutes, d=days, and y=years).

Heuristic search parameter	0.001	0.005	0.01	0.025	0.05	0.1	0.2	0.3	0.4	0.5
Detection rate	0.588	0.588	0.588	0.588	0.882	0.882	0.882	0.882	0.882	0.882
Running time	1.1m	2.5m	3.9m	9.1m	18.1m	**49.7m**	3.2h	11.9h	2.03d	6.44d
Heuristic search parameter	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.999	0.9999	1.0
Detection rate	0.882	0.882	0.882	0.882	0.882	0.882	0.882	0.882	0.882	0.882
Running time	23.3d	85.5d	308.9d	4.2y	**19y**	77y	390y	5007y	32224y	**363930y**

II.2.3.D Results on the identification of interaction effects among significant SNPs

Figure 19 gives the comparison of SCA and logistic regression with interaction terms on the identification of interaction effects among SNPs with significant marginal effects. On each of the five interaction models in the simulation, SCA consistently outputs a smaller p-value, which suggests a better power. The detailed p-values associated with each interaction model are presented in Table 25. For the interaction models with more than or equal to 3 SNPs, we also examine the interaction effect of sub-models that have strictly smaller size. For example, model 2 (which consists of A, B, and C) has three sub-models, [A, B], [A, C] and [B, C]. All the sub-models also have smaller p-values under our SCA test than logistic regression with interaction terms.

Figure 20 shows Q-Q plot for SCA tested under the null hypothesis on the interaction effects among SNPs with significant marginal effects. Each blue circle in the figure is a pair of SNPs that belong to different interaction models. These pairs have no interaction effects and should

follow the distribution under the null hypothesis. The distribution of the p-value under the null hypothesis is a uniform distribution. The red line is drawn according to the function "$y = x$". The blue circles are close to the red line, which supports that the smaller p-values shown in Figure 19 and Table 25 are not because of inflation of our calculated p-values and that the SCA test on the interaction effects among SNPs with significant marginal effects does follow the distribution specified by (14).

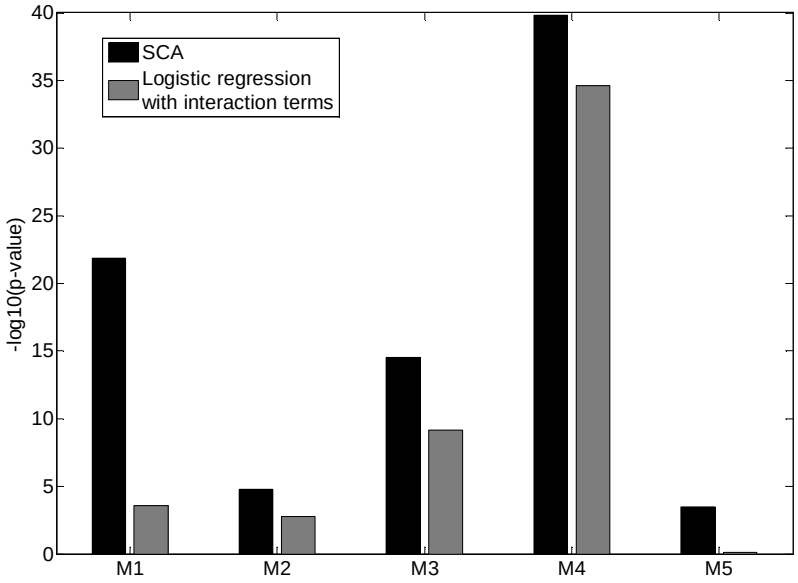

Figure 19: Comparison on detecting interaction effects among SNPs with significant marginal effects.

Table 25: The significance level for each interaction model and its sub-models

	Our proposed method	Logistic regression with interaction terms
Model 1	[A, B, C, D, E]: 1.3×10^{-22}	[A, B, C, D, E]: 2.5×10^{-4}
Model 2	[A, B]: 1.2×10^{-9} [A, C]: 3.6×10^{-9} [B, C]: 0.780 [A, B, C]: 1.6×10^{-5}	[A, B]: 7.1×10^{-3} [A, C]: 3.5×10^{-3} [B, C]: 0.778 [A, B, C]: 1.7×10^{-3}
Model 3	[A, B]: 8.2×10^{-6} [A, C]: 2.8×10^{-3} [B, C]: 5.2×10^{-5} [A, B, C]: 2.9×10^{-15}	[A, B]: 1.26×10^{-5} [A, C]: 5.8×10^{-3} [B, C]: 7.5×10^{-5} [A, B, C]: 6.7×10^{-10}
Model 4	[A, B]: 1.46×10^{-40}	[A, B]: 2.5×10^{-35}
Model 5	[A, B]: 3.2×10^{-4}	[A, B]: 0.756

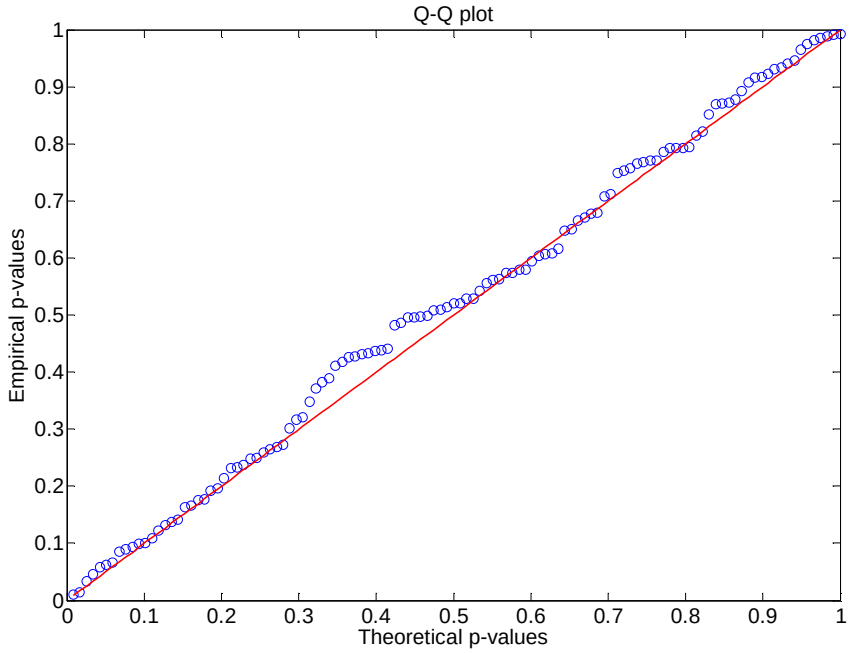

Figure 20: Q-Q plot for SCA test under the null hypothesis

II.2.4 Results on Real Datasets

We have applied our proposed approach to four real datasets: MESA, DHS, SLEGEN and Prostate Cancer. These datasets vary in terms of the SNP number, the sample size, and the diseases under study.

II.2.4.A MESA data

MESA stands for Multi-Ethnic Study of Atherosclerosis (i.e. hardening of the arties). There are 1572 subjects in total and 2659 SNPs genotyped. The subjects are classified into two groups, case and control, based on the calcium level. Each case is matched to two controls by age, race and gender. Thus, we have 524 cases and 1048 controls.

SCA-HCIG was applied to the MESA data and up to the fifth order interactions were searched. No single SNP or SNP subset passed the 0.05 significance threshold after Bonferroni correction. The top 5 interactions for each order are listed in Table 26.

Table 26: The top interactions discovered by SCA on MESA data

Marginal effect		Second order interaction		Third order interaction		Fourth order interaction	
rs#	p-val	rs#	p-val	rs#	p-val	rs#	p-val
rs8181474	0.000443	rs8181474 rs4987310	1.49E-05	rs11200607 rs1136159 rs1403543	3.47E-09	rs3791676 rs651821 rs673548 rs7947104	2.13E-12
rs12490383	0.000828	rs11088251 rs11716763	1.70E-05	rs8181474 rs1403543 rs1136159	8.70E-09	rs3791676 rs651821 rs1042034 rs7947104	1.45E-11
rs3753306	0.002213	rs8181474 rs1377175	1.91E-05	rs7076748 rs989692 rs8181474	1.02E-08	rs3791676 rs651821 rs676210 rs7947104	1.46E-11
rs290494	0.002617	rs8181474 rs11130129	1.93E-05	rs8181474 rs27184 rs7254487	1.15E-08	rs8181474 rs1403543 rs1136159 rs7918867	4.74E-11
rs11088251	0.003303	rs8181474 rs11716763	1.95E-05	rs3791676 rs10750096 rs585800	4.60E-08	rs8181474 rs1403543 rs1136159 rs13077421	8.07E-11

II.2.4.B DHS data

DHS is the Diabetes Heart Study from Dr. Bowden of Wake Forest University. There are 6 sets of binary traits: T2DM, MS, CVD1, CVD2, CAC400, and MEGAPHEN, so there are 6 datasets. The size for each dataset (subject size by SNP size) is:

T2DM: 1082 by 455

MS: 1082 by 455

CVD1: 1008 by 455

CVD2: 1024 by 455

CAC400: 914 by 455

MEGAPHEN: 863 by 455

We applied SCA on the 6 data sets to detect whether there is a significant interacting SNP subset. Interactions up to the fourth order were interrogated in the experiments. The top SNP subsets are listed in Table 27 when their p-values are below 20 times of 0.05 experimental-wise significant level. The significant associations with Bonferoni corrected p-values less than 0.1 are highlighted by bold and red. A fourth-order interaction is associated with CAC400, with the Bonferroni corrected p-value of 0.04.

Table 27: The top interactions discovered by SCA on DHS data. B-C p-value stands for the Bonferroni corrected p-value.

T2DM		MS		CVD1	
rs#	B-C p-value	rs#	B-C p-value	rs#	B-C p-value
ABCA1rs2230806	0.261138	None.		vdr	0.407187
ALOX12_7478	0.379724				
ALOX5rs1369214	0.564853				
CVD2		CAC400		MEGAPHEN	
rs#	B-C p-value	rs#	B-C p-value	rs#	B-C p-value
Chr16FM_rs210715 **Chr16FM_rs1433747** **ENPP1_rs1044498**	**0.090483**	**Chr14FM_rs2022703** **KALRN_rs11929003** **KCNE1_rs2070357** **alox15_1054**	**0.040141**	None.	

Chr14FM_rs715267 KCNE1_rs2070357 Chr14FM_rs1713430	0.206256	**ESR1rs9479097**	**0.052112**	
AHSGrs4917 APM1_G276T Chr14FM_rs1755784	0.261664	**ENPP1_rs1044498**	**0.077313**	
ALOX5AP_4431 Chr16FM_rs210715 Chr16FM_rs1433747 ENPP1_rs1044498	0.281912	Chr14FM_rs2022703	0.16311	
ENPP1_rs1044498	0.28848	Chr14FM_rs2022703 KCNE1_rs2070357 ephx2rs721619	0.357873	
Chr16FM_rs3928713 vdr	0.30197	Chr14FM_rs2022703 KALRN_rs1444746 KCNE1_rs2070357 alox15_1054	0.395979	
vdr	0.327219	KALRN_rs1444746 KCNE1_rs2070357 alox15_1054	0.403309	
ESR1rs9479097	0.358577	Chr16FM_rs11075183 Chr14FM_rs2022703	0.516834	
ESR1rs3778609	0.397846	ARRB2_rs9910421	0.534442	
		ENPP1_rs1044498 Chr16FM_rs150929	0.54556	

II.2.4.C SLEGEN data

The original study design of SLEGEN data consists of two independent data sets, LUPUS and LLAS. LUPUS is the dataset of top 8230 SNPs out of total 317,501 SNPs with Illumania SNP chip. 707 cases and 2318 controls are genotyped in LUPUS data. LLAS has genotyped 8230 SNPs which are extracted form LUPUS data, and it contains 1760 cases and 2083 controls.

To test our algorithm, we have used LLAS as the initial dataset to discover on and applied the LUPUS dataset as validation cohort. The reason we do not choose LUPUS as the discovery dataset is that LUPUS will make the threshold of the claimed significant findings difficult to define.

The detailed results are illustrated in Table 28, Table 29, and Table 30. Summarily,

(1), 64 one-way SNPs (Table 28) and 23 two-way interacting SNPs subsets (Table 29) are associated with the disease significantly through the analysis of the LLAS dataset; Among the 64 significant one-way SNPs, 30 significant pairs (Table 30) interactingly define the disease. The discoveries in Table 28 are detected by taking into account the interaction effects with the purpose of finding more markers associated with the disease; while the discoveries in Table 30 are picked up to indicate which SNPs are interacting with each other under the premise of known significant marginal effects.

(2), 60 out of 64 positives based on the one-way marginal effect are validated on the LUPUS dataset. If the p-value on the LUPUS dataset is larger than 0.05/38.58=0.0013, we consider it as not validated. The factor of 38.58 here comes from the fact that LUPUS is a deduced dataset of 8230 SNPs from the original 317,501 SNP dataset.

(3), 16 out of the 23 positives based on the two-way interaction are fully validated and 2 out of 23 are partially validated. Because these two-way discoveries are detected through the incorporation of interaction effects, we evaluate the interaction effect in the validation dataset

with the marginal effects removed. The discoveries with p-value larger than 0.1 are interpreted as not validated, and those with p-value between 0.05 and 0.1 are considered as partially validated.

(4), All the 30 significant two-way interacting SNP subsets with significant marginal effects are validated on the LUPUS dataset. All these discoveries have p-values less than 0.05.

Table 28: Findings with marginal effects by SCA on SLEGEN data.

SNP index	p-value after Bonferroni Correction	p-value on the independent cohort	Validated?	Notes
3001	1.41E-16	1.14E-20	Yes	
2995	2.63E-16	4.51E-20	Yes	
3006	1.54E-15	1.90E-18	Yes	
3020	1.44E-14	3.07E-18	Yes	
2986	4.02E-13	1.39E-15	Yes	
2977	3.35E-12	2.72E-13	Yes	
2970	1.35E-11	5.35E-16	Yes	
3021	1.91E-10	1.55E-16	Yes	
3019	1.13E-09	3.42E-13	Yes	
2964	1.51E-09	3.79E-10	Yes	
2963	1.70E-08	1.08E-11	Yes	
3013	4.83E-08	1.25E-11	Yes	
3024	5.69E-08	1.18E-13	Yes	
2950	1.47E-07	1.27E-09	Yes	
2953	2.01E-07	3.23E-11	Yes	
2955	2.64E-07	3.16E-09	Yes	
3015	8.89E-07	2.00E-14	Yes	
3010	1.17E-06	7.39E-11	Yes	
2933	2.56E-06	2.37E-08	Yes	
2944	4.43E-06	1.33E-09	Yes	
2943	4.66E-06	1.55E-10	Yes	
2921	6.55E-06	4.61E-08	Yes	
2971	8.74E-06	2.72E-13	Yes	
2973	1.41E-05	3.58E-04	Yes	
6733	1.62E-05	4.40E-09	Yes	
2925	1.73E-05	6.65E-09	Yes	

2993	1.94E-05	3.23E-15	Yes	
2987	2.01E-05	6.62E-10	Yes	
2930	2.08E-05	1.21E-07	Yes	
2932	2.80E-05	1.20E-08	Yes	
3781	3.24E-05	6.17E-10	Yes	
3025	3.54E-05	1.98E-06	Yes	
2916	5.55E-05	5.43E-08	Yes	
2978	8.72E-05	1.45E-05	Yes	
2982	0.000153349	2.34E-09	Yes	
3004	0.000184485	6.71E-06	Yes	
2965	0.00018501	5.62E-05	Yes	
3002	0.000203327	1.24E-05	Yes	
2966	0.00024513	3.50E-07	Yes	
2940	0.00033211	1.72E-09	Yes	
2911	0.000337934	2.53E-05	Yes	
2946	0.000530391	1.72E-06	Yes	
3005	0.000602349	5.19E-09	Yes	
3012	**0.000884123**	**2.08E-03**	**No**	**p-val on the independent cohort is larger than the threshold 0.0013**
3017	0.000885734	3.06E-10	Yes	
3008	0.0011048	9.46E-07	Yes	
2910	0.002316	6.05E-08	Yes	
3000	0.00266177	8.20E-13	Yes	
2909	0.00398238	3.03E-06	Yes	
2907	0.004082	6.46E-06	Yes	
2894	0.00491764	4.87E-08	Yes	
2976	**0.00718452**	**2.20E-02**	**No**	**p-val on the independent cohort is larger than the threshold 0.0013**
2901	0.0078933	4.71E-08	Yes	
2994	0.0085487	5.82E-06	Yes	
6734	0.00921906	1.27E-04	Yes	
2947	**0.0108479**	**2.44E-02**	**No**	**p-val on the independent cohort is larger than the threshold 0.0013**
2991	0.0112237	6.32E-09	Yes	
7713	**0.0114679**	**9.18E-02**	**No**	**p-val on the independent cohort is larger than the**

				threshold 0.0013
2999	0.0142218	8.86E-12	Yes	
3022	0.0144104	2.55E-09	Yes	
2998	0.0227439	4.13E-12	Yes	
2905	0.0271235	1.95E-04	Yes	
5088	0.0294004	4.48E-05	Yes	
6735	0.0415087	3.97E-04	Yes	

Table 29: Findings with 2-order Interaction by SCA on SLEGEN data.

SNP1 Index	SNP2 Index	p-value after Bonferroni Correction	p-value on the independent cohort	Validated	Notes
2985	3011	1.01E-05	6.30E-05	Yes	
2968	2981	3.61E-05	9.15E-05	Yes	
2997	3011	4.77E-05	2.28E-05	Yes	
2968	2985	0.00161719	0.0965	Partially	the pval on the independent cohort is in the interval between 0.05 and 0.1
2996	3011	0.00219205	1.84E-06	Yes	
2975	3011	0.00299981	7.70E-03	Yes	
2982*	3011	0.00406481	1.50E-03	Yes	
2962	2985	0.00424927	5.00E-02	Yes	
2974	3011	0.00538608	7.80E-03	Yes	
2937	2962	0.00603593	6.30E-05	Yes	
2976*	3011	0.00747255	0.099	Partially	the pval on the independent cohort is in the interval between 0.05 and 0.1
2974	2981	0.0132561	2.03E-02	Yes	
3011	3913	0.0156054	0.2435	No	the pval on the independent cohort is larger than 0.1
3012*	1102	0.0167862	0.9315	No	the pval on the independent cohort is larger than 0.1
2985	3018	0.0338061	8.40E-03	Yes	
1102	8079	0.0338255	0.9797	No	the pval on the independent cohort is larger than 0.1
3002*	3016	0.0433171	3.40E-06	Yes	
3010*	2996	0.0437915	1.10E-05	Yes	
3008*	2996	0.0453849	3.03E-07	Yes	
2976*	2968	0.0461939	6.30E-08	Yes	

2978*	2985	0.0488885	9.70E-08	Yes	
2905*	**8213**	**0.0103914**	**0.1131**	**No**	**the pval on the independent cohort is larger than 0.1**
2911*	**8213**	**0.0414658**	**0.4119**	**No**	**the pval on the independent cohort is larger than 0.1**

Table 30: Interactions among SNPs identified by SCA on SLEGEN data.

SNP1 Index	SNP2 Index	p-value after Bonferroni Correction	p-value on the independent cohort	Validated
2987	2993	1.64E-05	1.85E-03	Yes
2978	2982	0.000287686	1.73E-07	Yes
2993	3002	0.000870548	2.40E-03	Yes
2982	3005	0.00122941	1.49E-04	Yes
2991	3002	0.00128719	1.63E-06	Yes
2978	2987	0.00152522	3.35E-07	Yes
2978	2991	0.00254837	1.40E-05	Yes
2987	3013	0.00278143	5.94E-05	Yes
2976	2991	0.00293594	3.29E-03	Yes
2976	2982	0.00574539	2.27E-05	Yes
2978	2993	0.00946235	3.52E-04	Yes
3002	3013	0.00972175	1.94E-07	Yes
2946	2978	0.0114774	3.57E-07	Yes
2971	3005	0.0117917	1.06E-03	Yes
2987	3010	0.0122932	8.00E-07	Yes
2973	3005	0.0130887	1.65E-03	Yes
3000	3013	0.0143879	1.43E-05	Yes
2993	3000	0.0169381	1.57E-04	Yes
2976	2993	0.0219936	1.15E-02	Yes
2946	2973	0.0237444	5.55E-05	Yes
2946	2971	0.0284158	6.90E-05	Yes
2976	3005	0.0285831	1.91E-02	Yes
3000	3008	0.0302911	5.75E-05	Yes
3008	3013	0.0314413	2.62E-05	Yes
2987	3008	0.0351054	9.38E-06	Yes
2982	3010	0.0358266	1.02E-04	Yes
2982	3000	0.0369303	6.22E-04	Yes
2987	3000	0.0384559	1.10E-04	Yes
2946	2976	0.04027	1.87E-02	Yes
2987	3005	0.0484218	3.42E-06	Yes

Due to the LD structure, we desired to identify a single SNP to represent each region. This was accomplished via stepwise logistic regression (Harley, et al., 2008). Table 31 shows the six representative SNPs from six genes/regions that are significantly associated with the SLE with marginal effects. Each pair out of the six regions, totally 15 pairs, is tested for the interaction effects. The p-values associated with all the pairs are provided in Table 32. Under the 0.05 significance level, the experimental-wise significance threshold is 0.0033. Among all 15 pairs, two pairs pass the threshold and are considered to be significant. The first pair (IRF5/TNPO3-rs12537284 and HLA region 1 - rs3131379) has a p-value of 8.05×10^{-6}, which is 1.20×10^{-4} after Bonferroni correction. The second pair (BLK/c80rf12-rs7836059 and HLA region 2 - rs9275572) has a p-value of 5.51×10^{-5}, which is 8.26×10^{-4} after Bonferroni correction.

Both two pairs are consistent in the direction across two datasets LLAS and LUPUS, in terms that the combination of risk alleles increases disease risk for both two datasets. And both pairs have p-values less than 0.05 if the test is implemented separately on the two datasets. Figure 21 demonstrates the proportion of cases in each genotype specified by the SNP pair - IRF5/TNPO3 - rs12537284 and HLA region 1 - rs3131379. Each SNP is dichotomized through dominant/recessive genetic coding. The left subfigure shows the disease distribution for the LLAS dataset and the right subfigure shows the disease distribution for the LUPUS dataset. If tested separately on each dataset, the p-values are 0.0025 and 0.00024, respectively. Table 33 shows the detailed subject distribution for each genotype specified by the SNP pair - IRF5/TNPO3-rs12537284 and HLA region 1 - rs3131379. In each cell, the first number is the number of disease subjects and the second number is the total number of subjects carrying this genotype. The bold font indicates the genotype that potentially possesses the interaction.

Figure 22 presents the proportion of cases in each genotype specified by the SNP pair - BLK/c80rf12 - rs7836059 and HLA region 2 - rs9275572. Each SNP is dichotomized through dominant/recessive genetic coding. The left subfigure shows the disease distribution for the LLAS dataset and the right subfigure shows the disease distribution for the LUPUS dataset. If tested separately on each dataset, the p-values are 0.015 and 0.0002, respectively. Table 34 illustrates the detailed subject distribution for each genotype specified by the SNP pair - BLK/c80rf12-rs7836059 and HLA region 2 - rs9275572. In each cell, the first number is the number of disease subjects and the second number is the total number of subjects carrying this genotype. The bold font indicates the genotypes that potentially possess the interaction.

Table 31: Six genes/regions are significantly associated with SLE with marginal effects

Gene	HLA region 1	HLA region 2	IRF5/TNPO3	BLK/c80rf12	KIAA1542	ITGAM
rs#	rs3131379	rs9275572	rs12537284	rs7836059	rs4963128	rs9888739
chr.	6p21.33	6p21.32	7q32.1	8p23.1	11p15.5	16p11.2

Table 32: The p-values associated with all pair combinations in the six regions. Two pairs pass the experimental-wise significance threshold 0.0033, which are highlighted by the bold font.

rs#	rs3131379	rs9275572	rs12537284	rs7836059	rs4963128	rs9888739
rs3131379						
rs9275572	0.0845					
rs12537284	***5.51x10^{-5}***	0.1242				
rs7836059	0.1582	***8.05 x10^{-6}***	0.5991			
rs4963128	0.0651	0.3273	0.0379	0.1523		
rs9888739	0.4321	0.1555	0.1570	0.3926	0.5652	

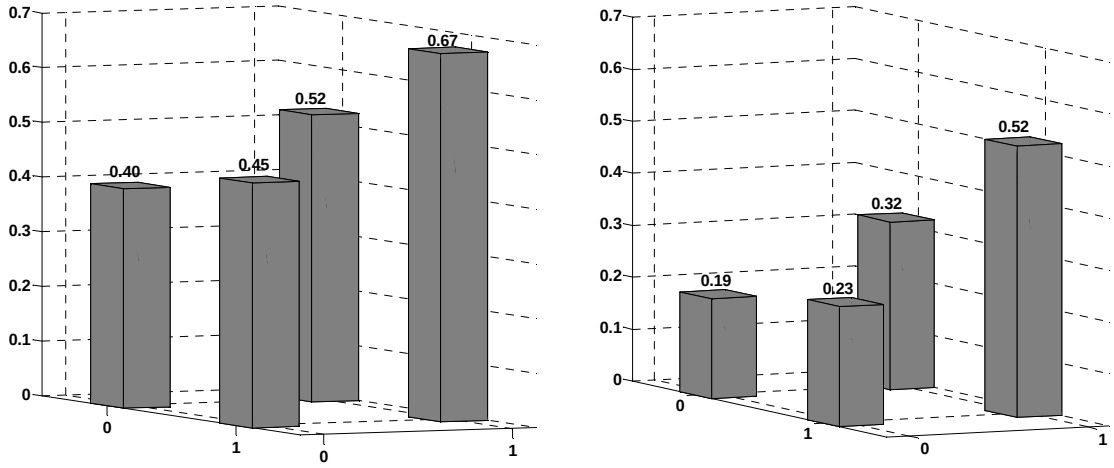

Figure 21: The proportion of cases in each genotype specified by the SNP pair - IRF5/TNPO3-rs12537284 and HLA region 1 - rs3131379. Each SNP is dichotomized through dominant/recessive genetic coding. The left subfigure shows the disease distribution for the LLAS dataset and the right subfigure shows the disease distribution for the LUPUS dataset. If tested separately on each dataset, the p-values are 0.0025 and 0.00024, respectively.

Table 33: The detailed subject distribution for each genotype specified by the SNP pair - IRF5/TNPO3-rs12537284 and HLA region 1 - rs3131379. In each cell, the first number is the number of disease subjects and the second number is the total number of subjects carrying this genotype. The bold font indicates the genotype that potentially possesses the interaction.

LLAS dataset (rs12537284 and rs3131379)				LUPUS dataset (rs12537284 and rs3131379)			
	AA	AB	BB		AA	AB	BB
AA	90 / 240	191 / 551	134 / 308	AA	40 / 247	73 / 483	52 / 267
AB	159 / 406	410 / 921	244 / 536	AB	79 / 354	151 / 702	98 / 383
BB	76 / 158	222 / 409	**159 / 236**	BB	38 / 116	93 / 290	**82 / 157**

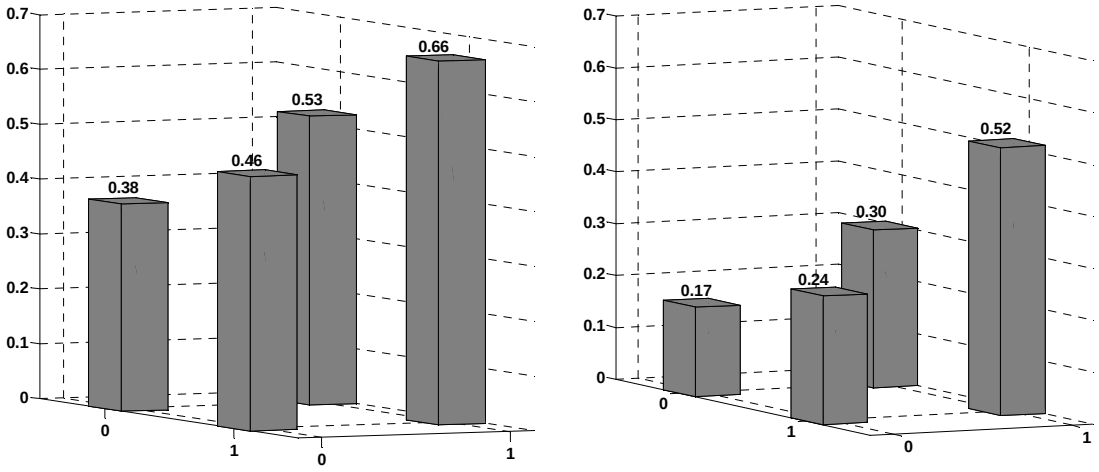

Figure 22: The proportion of cases in each genotype specified by the SNP pair - BLK/c80rf12-rs7836059 and HLA region 2 - rs9275572. Each SNP is dichotomized through dominant/recessive genetic coding. The left subfigure shows the disease distribution for the LLAS dataset and the right subfigure shows the disease distribution for the LUPUS dataset. If tested separately on each dataset, the p-values are 0.015 and 0.0002, respectively.

Table 34: The detailed subject distribution for each genotype specified by the SNP pair - BLK/c80rf12-rs7836059 and HLA region 2 - rs9275572. In each cell, the first number is the number of disease subjects and the second number is the total number of subjects carrying this genotype. The bold font indicates the genotypes that potentially possess the interaction.

LLAS dataset (rs7836059 and rs9275572)				LUPUS dataset (rs7836059 and rs9275572)			
	AA	AB	BB		AA	AB	BB
AA	747 / 1980	317 / 671	25 / 56	AA	291 / 1690	140 / 572	14 / 50
AB	347 / 667	**183 / 282**	**18 / 24**	AB	128 / 440	**85 / 282**	**13 / 17**
BB	33 / 55	**17 / 22**	**2 / 4**	BB	20 / 47	**9 / 15**	**2 / 2**

II.2.4.D Prostate Cancer 16 SNPs

16 SNPs are selected based on a survey of the literatures, through which significant associations with prostate cancers were reported. An independent dataset based on a Swedish population was

genotyped, of which there are 2893 subjects with prostate cancer and 1781 control subjects. The purpose of this study is to see whether there are interactions among these SNPs as these SNPs are significant and well-established markers for prostate cancer.

Applying the SCA algorithm (more specifically, the second function of identifying interaction effects between SNPs with significant marginal effects), we have found a statistically significant three-way interaction. Three interacting SNPs [rs721048, rs10486567 and rs1447295] have the p-value 3.02e-006, which is experimental-wise significant (0.0017 after multiple test correction). It seems to be a true interaction, because it is partially verified with the public CGEMS data. rs721048 is not genotyped in the CGEMS data. Based on the CGEMS data, the interaction between rs10486567 and rs1447295 is also significant, with the p-value 0.018.

II.2.4.E Interaction between thrombophilic mutations and oral contraceptive on the venous thrombosis

The interaction of thrombophilic mutations with oral contraceptives on venous thrombosis is a pronounced example of a gene-environment interrelationship study. A venous thrombosis is a blood clot that forms within a vein, especially in the deep veins of the legs or in the pelvic veins. There are both genetic and environmental risk factors. The R506Q mutation of factor V and the G20210A mutation of prothrombin are two thrombophilic genetic factors (Rosendaal, 1999), moderately common in whites with frequencies of 5% and 2%, respectively. Factor V is a protein of the coagulation system and factor Va (activated factor V) is a highly procoagulant cofactor in the generation of thrombin, which is a crucial element in blood clotting. The R506Q substitution in factor V involves one of three sites that are cleaved by activated protein C. This

mutation slows down the proteolytic inactivation of factor Va, which in turn leads to the augmented generation of thrombin (Seligsohn and Lubetsky, 2001). Prothrombin is proteolytically cleaved to form thrombin. The G20210A mutation in the 3' untranslated region of the prothrombin gene is associated with an increased level of plasma prothrombin, promoting the generation of thrombin and impairing the inactivation of factor Va by activated protein C (Seligsohn and Lubetsky, 2001). The use of the oral contraceptive has long been recognized as a risk factor for venous thrombosis. The oral contraceptive has significant effect on the generation of thrombin, by both decreasing the level of factor V and increasing the level of prothrombin.

The interaction between thrombophilic mutations and oral contraceptive is well established with multiple epidemiological and mechanical studies (Legnani, et al., 2002; Martinelli, et al., 1999; Rosing, et al., 1999; Vandenbroucke, et al., 1994). Table 35 and Table 36 show two studies illustrating the interaction between the thrombophilic genetic mutation and the use of oral contraceptive. In Legnani et al.'s study, the odds ratio associated with the use of oral contraceptive but no thrombophilic genetic risk mutation is 1.95, and the odds ratio associated with genetic defects but no use of contraceptive is 4.79. According to the multiplicative model, the odds ratio associated with the presence of both risk factors should be 9.34, while the observed odds ratio is 27.4. This is a strong evidence of interaction. Indeed, by applying the logistic regression model, we would get a p-value of 0.021, which is statistically significant. If we apply the new proposed model, we have a p-value of 0.00062. There are 947 subjects in Legnani et al.'s study. When all the frequencies of the risk factors and the effect size are kept the same, we estimate that, to achieve the 0.05 significance level, the logistic regression model

would require 676 subjects, while the new model needs only 303 subjects. We see that the sample size required for log-regression model is significantly reduced.

For the Martinelli et al.'s study, the odds ratio associated with the presence of both risk factors (according to a multiplicative model) is expected to be 11.9, compared to the observed value of 18.1. Both studies have the same effect direction, that is, the observed odds ratio is larger than the expectation. Due to the limited sample size, the conclusion is not statistically significant in the Martinelli et al.'s study. The p-value generated by the logistic regression model is 0.618 and the p-value obtained from the new model is 0.183. To achieve the 0.05 significance level, the estimated sample size associated with the logistic regression model is 4,391, while we would only require 614 subjects for the new model. In this case, the new model only needs one seventh of the sample size necessary for the logistic regression model.

Table 35: Legnani et al.'s study: risk of venous thrombosis according to the presence of thrombophilic genetic mutation and the use of oral contraceptive.

thrombophilic genetic risk mutation	oral contraceptive	controls	cases	odds ratio
-	-	444	118	1
-	+	166	86	1.95
+	-	33	42	4.79
+	+	7	51	27.4

Table 36: Martinelli et al.'s study: risk of venous thrombosis according to the presence of thrombophilic genetic mutation and the use of oral contraceptive.

thrombophilic genetic risk mutation	oral contraceptive	controls	cases	odds ratio
-	-	127	35	1
-	+	41	52	4.60
+	-	7	5	2.59
+	+	4	20	18.1

II.2.4.F Interaction between NAT2 gene and smoking on bladder cancer

With hundreds of thousands of new cases diagnosed each year worldwide, bladder cancer is increasingly important for public health. Tobacco smoking is the predominant known risk factor for bladder cancer. In Europe, smoking is estimated to cause over half of bladder cancer cases in men and one-third of cases among women (Zeegers, et al., 2000). Multiple carcinogens have been found in tobacco smoke, including polycyclic aromatic hydrocarbons, N-nitrosamines, aromatic amines, heterocyclic amines, and aldehydes. Originally inert, these carcinogens may undergo both activation and detoxification. Imbalance between activation and detoxification will increase the bladder cancer risk through accumulation of active arcinogen metabolites and increased DNA adduct formation (Gu, et al., 2005).

The NAT2 gene encodes an enzyme that functions to both activate and deactivate arylamine and hydrazine drugs and carcinogens (Sanderson, et al., 2007). The NAT2 enzyme is particularly active in the liver, gastrointestinal tract, and urinary bladder, among other organs and tissues. Due to the metabolic rate of exogenous compounds, the polymorphisms in the NAT2 gene can be classified into two types, rapid acetylator and slow acetylator. NAT2 slow acetylator is very

common in the Caucasian population, estimated to be around 55%. The association of the NAT2 slow acetylator with bladder risk is quite well established, serving as an outstanding example prior to the GWAS era for the replicated association between common genetic polymorphisms and complex diseases.

Multiple studies have consistently shown the interaction between the NAT2 gene and smoking on bladder cancer(Garcia-Closas, et al., 2005; Gu, et al., 2005; Sanderson, et al., 2007). Table 37 presents the non-meta-analysis study with the largest sample size (Garcia-Closas, et al., 2005). Choosing the bladder cancer risk at never smoking and NAT2 fast acetylator as the reference, the odds ratio associated with ``ever smoking'' (*i.e.*, an individual who has smoked before) and NAT2 fast acetylator is 1.86, and the odds ratio associated with never smoking and NAT2 slow acetylator is 0.91. According to the multiplicative model, the odds ratio associated with the presence of both risk factors should be 1.69, while the observed odds ratio is 2.89. So the interaction is evident. Indeed, by applying the logistic regression model, we would get a p-value of 0.015, which is statistically significant. When we apply the new proposed model, we get a p-value of 0.0011. There are totally 2,264 subjects in the study. When all the frequencies of the risk factors and the effect size are kept the same, we estimate that, to achieve the 0.05 significance level, the logistic regression model would require 1,449 subjects and the log-regression model needs only 796 subjects. We see that the sample size required for log-regression model is significantly reduced.

This interaction has strong biological plausibility since aromatic amines (e.g., 4-aminobiphenyl) and heterocyclic amines (e.g., PhIP) are two of the primary carcinogens found in tobacco smoke,

and NAT2 enzyme has been identified as being involved in the metabolism of these carcinogens via N-, O-, or N,O-acetylation. Generally speaking, N-acetylation is a detoxification step, and O-acetylation is an activation step. NAT2 slow acetylator has a decreased capacity to detoxify aromatic monoamines by N-acetylation. The individual with NAT2 slow acetylation genotype and smoking habit will accumulate the active carcinogens and hence increase the risk of bladder cancer.

Table 37: Joint association for tobacco smoking status and NAT2 acetylation genotype with bladder cancer risk.

NAT2 acetylation genotype	smoking status	controls	cases	odds ratio
fast	never	131	66	1
fast	ever	362	340	1.86
slow	never	199	91	0.91
slow	ever	438	637	2.89

II.2.4.G Interaction between ALDH2 gene and alcohol consumption on esophageal cancer

Both ALDH2 and alcohol consumption are known genetic/environmental factors that are associated with esophageal cancer. Heavy alcohol drinking has been found to be a risk factor to esophageal cancer in many epidemiological studies (Allen, et al., 2009). When alcohol is metabolized in liver, it is broken to acetaldehyde, which is oxidative and recognized as a carcinogen by binding to cellular protein and DNA. The majority (99%) of the produced acetaldehyde is eliminated by the liver. The ALDH2 protein is responsible for degrading the remaining. There is a functional polymorphism in the ALDH2 gene, namely ALDH2 Glu478Lys. The Glu allele encodes a protein with normal catalytic activity, while the Lys allele encodes an

inactive protein. A defect in the ALDH2 genes significantly reduces the capacity to degrade acetaldehyde and hence exposes an individual to more acetaldehyde than normal. It is biologically plausible for the ALDH2 protein and alcohol consumption to interactingly influence the risk of esophageal cancer (Lewis and Smith, 2005; Matsuo, et al., 2001).

Figure 23 shows the re-analysis of the interaction between ALDH2 gene and alcohol consumption. The data is collected from the first study of the ALDH2-alcohol interaction effect on esophageal cancer. The original report discovered the interaction effect through logistic regression analysis, which has been confirmed by follow-up studies (Lewis and Smith, 2005) and indicated that it was a true interaction. The distribution of the cases and the controls are presented in the figure. We re-analyze the data using our proposed model. The significance through our model is 7.4e-6, compared to a p-value of 2.5e-3 with the logistic regression model. The p-value generated from our method is much smaller than the conventional one. It is an improvement of almost 3 orders. There are in total 343 subjects in the study. When all the frequencies of the risk factors and the effect size are kept the same, we estimate that, to achieve the 0.05 significance level, the logistic regression model would require 142 subjects and the new model needs only 64 subjects. We see that the sample size required for the new model is significantly reduced.

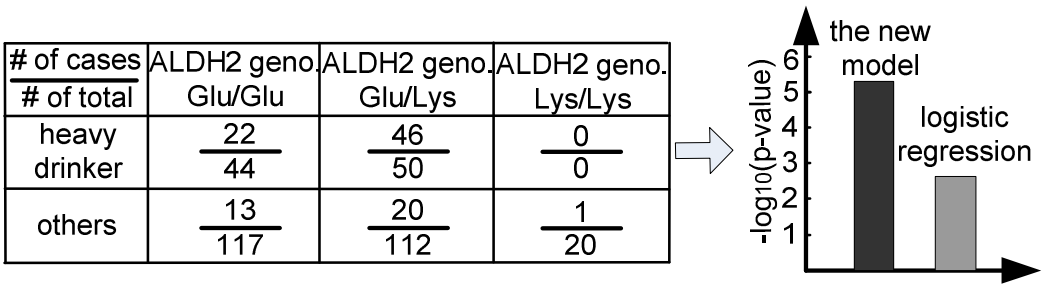

Figure 23: Re-analysis of the interaction between gene ALDH2 and alcohol consumption.

II.2.4.H Interaction between tobacco smoking and alcohol drinking on esophageal cancer

It has long been suggested that tobacco smoking and alcohol drinking interplay to influence the risk of cancer (Garro and Lieber, 1990). Alcohol may act as a cocarcinogen and enhance the carcinogenic effects of other chemicals from tobacco smoking. Indeed, quite a few epidemiological studies have confirmed their interaction effect on esophageal cancer (Castellsague, et al., 1999; Lee, et al., 2005).

Castellsague et al.'s report (Castellsague, et al., 1999) is probably the first large scale case-control study implying the interaction effect of tobacco smoking and alcohol drinking on esophageal cancer. The study showed that the combination of the two factors significantly increased the disease risk more than either of them separately. However, although the report demonstrated the statistical evidence on both groups of males and all subjects, it failed to see the significant interaction on the males group. By applying the new model, the new analysis generates consistent results.

Table 38 presents the subjects distribution specified by the status of alcohol drinking, tobacco smoking and esophageal cancer in Castellsague et al.'s study. The data are divided into three groups, males, females and all. In each group, we calculate the interaction effect based on the logistic regression model and the new model. We can see that the new model consistently generates smaller p-values than the logistic regression model. In the males group, the p-value is 5.43e-6 based on the new model, while it is 0.81 for the logistic regression model and far from

being considered as significant. We also estimate the sample sizes required for both two models to achieve the 0.05 significance level, assuming that all the frequencies of the risk factors and the effect size are kept the same. In the males group, the logistic regression model needs 131,413 subjects, compared to 374 subjects required for the new model. In the females group, it is 339 for the logistic regression model and 235 for the new model. In the 'all' group, 596 subjects are necessary for the logistic regression model, while 312 subjects are enough for the new model.

Table 38: Joint association for alcohol drinking status and tobacco smoking status with esophageal cancer risk.

alcohol	smoking	males			females			all		
		controls	cases	odds ratio	controls	cases	odds ratio	controls	cases	odds ratio
never	never	189	8	1	234	83	1	423	91	1
never	ever	298	61	4.84	55	27	1.38	353	88	1.16
ever	never	144	24	3.94	63	29	1.30	207	53	1.19
ever	ever	777	562	17.1	19	36	5.34	796	598	3.49
Logistic regression (*p*)			0.81			0.014			5.10e-5	
Log regression (*p*)			5.43e-6			0.0031			2.11e-8	

III. PUG-OVRSVM to Select Multi-class Relevant Genes

III.1 Methods and Theory

In this section, we first discuss multicategory classification and associated feature selection, with an emphasis on OVRSVM and application to gene selection for the microarray domain. This discussion then naturally leads to our proposed PUG-OVRSVM scheme.

III.1.1 Maximum A Posteriori Decision Rule

Classification of heterogeneous diseases using gene expression data can be considered a Bayesian hypothesis testing problem (Hastie, et al., 2001). Let $\mathbf{x}_i = \left[x_{i1}, ..., x_{ij}, ..., x_{id} \right]$ be the real-valued gene expression profile associated with sample i across d genes for $i = 1, ..., N$ and $j = 1, ..., d$. Assume that the sample points $\mathbf{x}_i$ come from M classes, and denote the class conditional probability density function and class prior probability by $p(\mathbf{x}_i \mid \omega_k)$ and $P(\omega_k)$, respectively, for $k = 1, 2, ..., M$. To minimize the Bayes risk averaged over all classes, the optimum classifier uses the well-known maximum *a* posteriori (MAP) decision rule (Hastie, et al., 2001). Based on Bayes' rule, the class posterior probability for a given sample $\mathbf{x}_i$ is

$$P(\omega_k \mid \mathbf{x}_i) = \frac{P(\omega_k)\, p(\mathbf{x}_i \mid \omega_k)}{\sum_{k'=1}^{M} P(\omega_{k'})\, p(\mathbf{x}_i \mid \omega_{k'})} \tag{62}$$

and is used to (MAP) classify $\mathbf{x}_i$ to ω_k when

$$P(\omega_k \mid \mathbf{x}_i) > P(\omega_l \mid \mathbf{x}_i) \tag{63}$$

for all $l \neq k$.

III.1.2 Supervised Learning and Committee Classifiers

Practically, multicategory classification using the MAP decision rule can be approximated using parameterized discriminant functions that are trained by supervised learning. Let $f_k(\mathbf{x}_i, \boldsymbol{\theta})$, $k = 1, 2, ...M$, be the M outputs of a machine classifier designed to discriminate between M classes (>2), where $\boldsymbol{\theta}$ represents the set of parameters that fully specify the classifier, and with the output values assumed to be in the range [0,1]. The desired output of the classifier will be "1"

for the class to which the sample belongs and "0" for all other classes. Suppose that the classifier parameters are selected based on a training set so as to minimize the mean squared error (MSE) between the outputs of the classifier and the desired (class target) outputs,

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{M}\sum_{\mathbf{x}_i \in \omega_k}\left[\left[f_k(\mathbf{x}_i,\boldsymbol{\theta})-1\right]^2 + \sum_{l\neq k} f_l^{\,2}(\mathbf{x}_i,\boldsymbol{\theta})\right]. \tag{64}$$

Then, it can be shown that the classifier is being trained to approximate the posterior probability for class ω_k given the observed $\mathbf{x}_i$, i.e., the classifier outputs will converge to the true posterior class probabilities

$$f_k(\mathbf{x}_i,\boldsymbol{\theta}) \to P(\omega_k \mid \mathbf{x}_i) \tag{65}$$

if we allow the classifier to be arbitrarily complex and if N is made sufficiently large. This result is valid for any classifier trained with the MSE criterion, where the parameters of the classifier are adjusted to simultaneously approximate M discriminant functions, $f_k(\mathbf{x}_i,\boldsymbol{\theta})$ (Gish, 1990).

While there are numerous machine classifiers that can be used to implement the MAP decision rule (63) (Hastie, et al., 2001), a simple yet elegant way of discriminating between M classes, and which we adopt here, is based on an OVRSVM committee classifier (Ramaswamy, et al., 2001; Rifkin and Klautau, 2002; Statnikov, et al., 2005). Intuitively, each term within the sum over k in (64) corresponds to an OVR binary classification problem and can be effectively minimized by suitable training of a binary classifier (discriminating class k from all other classes). By separately minimizing the MSE associated with each term in (64) via binary classifier training and, thus, effectively minimizing the total MSE, a set of discriminant functions

$\left\{ f_k\left(\mathbf{x}_i, \boldsymbol{\theta}_k \subseteq \boldsymbol{\theta}\right)\right\}$ can be constructed which, given a new sample point, apply the decision rule (63), but with $f_k\left(\mathbf{x}_i, \boldsymbol{\theta}\right)$ playing the role of the posterior probability.

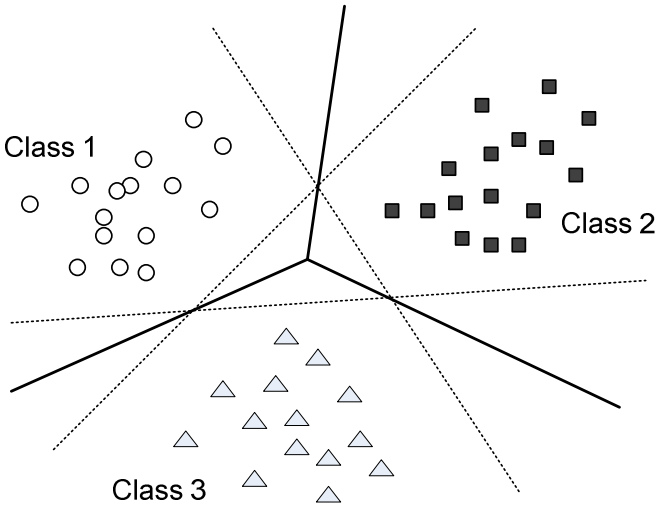

Figure 24: Conceptual illustration of OVR committee classifier for multicategory classification (three classes, in this case). The dotted lines are the decision hyperplanes associated with each of the component binary SVMs and the bold line-set represents the final decision boundary after the winner-take-all classification rule is applied.

Among the great variety of binary classifiers that use regularization to control the capacity of the function spaces they operate in, the best known example is the SVM (Hastie, et al., 2001; Vapnik, 1998). To carry over the advantages of regularization approaches for binary classification tasks to multicategory classification, the OVRSVM committee classifier uses M different SVM binary classifiers, each one separately trained to distinguish the samples in a single class from the samples in all remaining classes. For classifying a new sample point, the M SVMs are run, and the SVM that produces the largest (most positive) output value is chosen as the "winner" (Ramaswamy, et al., 2001). For more detailed discussion, see the critical review and experimental comparison by Rifkin and Klautau (Rifkin and Klautau, 2002). Figure 24 shows an

illustrative OVRSVM committee classifier for three classes. The OVRSVM committee classifier has proved highly successful at multicategory classification tasks involving finite or limited amounts of high dimensional data in real-world applications. OVRSVM produces results that are often at least as accurate as other more complicated methods including single machine multicategory schemes (Statnikov, et al., 2005). Perhaps more importantly for our purposes, the OVR scheme can be matched with an OVE gene selection method, as we elaborate next.

III.1.3 One-versus-everyone Fold-change Gene Selection

While gene selection is vital for achieving good generalization performance (Guyon, et al., 2002; Statnikov, et al., 2005), perhaps even more importantly, the identified genes, if statistically reproducible and biologically plausible, are "markers", carrying information about the disease phenotype (Wang, et al., 2008). We will propose two novel, effective gene selection methods for multicategory classification that are well-matched to OVRSVM committee classifiers, namely, OVR and OVE fold-change analyses.

OVR fold-change based PUG selection follows directly from the OVRSVM scheme. Let N_k be the number of sample points belonging to phenotype k; the geometric mean of the expression levels (on the untransformed scale) for gene j under phenotype k is

$$\mu_j(k) = \sqrt[N_k]{\prod_{i \in \omega_k} x_{ij}}\ , \tag{66}$$

$j = 1,...,d$; $k = 1,2,...,M$. Then, we define the OVRPUGs as:

$$\mathbb{J}_{\text{PUG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PUG}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \frac{\mu_j(k)}{\sqrt[M-1]{\prod_{l \neq k} \mu_j(l)}} \geq \tau_k \right. \right\}, \tag{67}$$

where $\{\tau_k\}$ are pre-defined thresholds chosen so as to select a fixed (equal) number of PUGs for each phenotype k. This PUG selection scheme (67) is similar to what has been previously proposed by (Shedden, et al., 2003):

$$\mathbb{J}_{\text{PUG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PUG}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \frac{\mu_j(k)}{\sqrt[N-N_k]{\prod_{i \notin \omega_k} x_{ij}}} \geq \tau_k \right. \right\}. \tag{68}$$

The critical difference between (67) and (68) is that the denominator term in (67) is the overall geometric center of the "geometric centers" associated with each of the remaining phenotypes while the denominator term in (38) is the geometric center of all sample points belonging to the remaining phenotypes. When $\{N_k\}$ are significantly imbalanced for different k, the denominator term in (68) will be biased toward the dominant phenotype(s).

However, a problem associated with both PUG selection schemes specified by (67) and (68) (and with the OVRSNR criterion (Golub et al., 1999)) is that the criterion function considers the remaining classes as a single super class, which is suboptimal because it ignores a gene's ability to discriminate between classes *within* the super class.

We therefore propose OVE fold-change based PUG selection to fully support the objective of multicategory classification. Specifically, the OVEPUGs are defined as:

$$\mathbb{J}_{\text{PUG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{PUG}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \frac{\mu_j(k)}{\max_{l \neq k} \left\{ \mu_j(l) \right\}} \geq \tau_k \right. \right\} \tag{69}$$

where the denominator term is the maximum phenotypic mean expression level over the remaining phenotype classes. This seemingly technical modification turns out to have important consequences since it assures that the selected PUGs are highly expressed in one phenotype relative to *each* of the remaining phenotypes, i.e. "high" (up-regulated) in phenotype k and "low" (down-regulated) in *all* phenotypes $l \neq k$. In our experimental results, we will demonstrate that (69) leads to better classification accuracy than (68) on a well-known multi-class cancer domain.

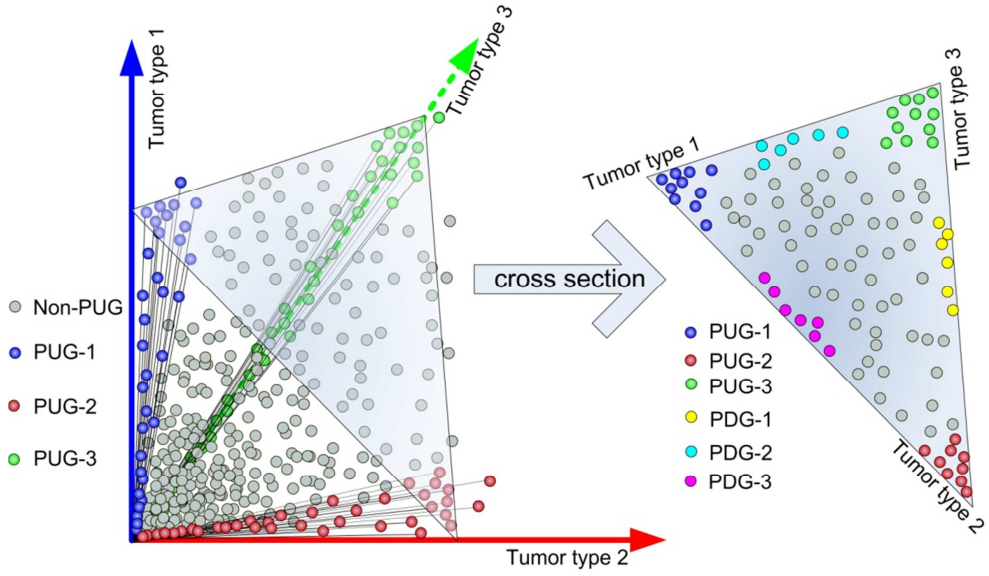

Figure 25: Geometric illustration of the selected one-versus-everyone phenotypic up-regulated genes (OVEPUGs) associated with three phenotypic classes. Three-dimensional geometric distribution (on the untransformed scale) of the selected OVEPUGs, which reside around the lateral-edges of the phenotypic gene expression scatter plot convex pyramid. A projected distribution of the selected OVEPUGs together with OVEPDGs is shown in the right cross-sectional plot, where OVEPDGs reside along the face-edges of the cross-sectional triangle.

Adopting the same strategy as in (Shedden, et al., 2003), to assure even-handed gene resources for discriminating both neighboring and well-separated classes, we select a fixed (common) number of top-ranked phenotype-specific subPUGs for each phenotype, i.e. $\left\|\mathbb{J}_{\mathrm{PUG}}(k)\right\| = N_{\mathrm{subPUG}}$ for all k, and pool all these subPUGs together to form the final gene marker subset $\mathbb{J}_{\mathrm{PUG}}$ for the OVRSVM committee classifier. In our experiments, the optimum number of PUGs per phenotype, N_{subPUG}, is determined by surveying the curve of classification accuracy versus N_{subPUG} and selecting the number that achieves the best classification performance. More generally, in practice, N_{subPUG} can be chosen via a cross validation procedure. Figure 25 shows the geometric distribution of the selected PUGs specified by (69), where the PUGs (highlighted data points) constitute the lateral-edge points of the convex pyramid defined by the scatter plot of the phenotypic mean expressions (Zhang, et al., 2008). Different from the PUG selection schemes given by (67) or (68), the PUGs selected based on (69) are most compact yet informative, since the down-regulated genes that are not differentially expressed between the remaining phenotypes (the genes on the lateral faces of the scatter plot convex pyramid) are excluded. From a statistical point of view, extensive studies on the normalized scatter plot of microarray gene expression data by many groups including our own indicate that the PUGs selected by (69) approximately follow an independent multivariate super-gaussian distribution (Zhao, et al., 2005) where subPUGs are mutually exclusive and phenotypic gene expression patterns defined over the PUGs are statistically independent (Wang, et al., 2003).

It is worth noting that the PUG selection by (69) also adopts a univariate fold-change evaluation that does not require calculation of either expression variance or of correlation between genes (Shi, et al., 2008). For the small sample size case typical of microarray data, multivariate gene

selection schemes may introduce additional uncertainty in estimating the correlation structure (Lai, et al., 2006; Shedden, et al., 2003) and thus may fail to identify true gene markers (Wang, et al., 2008). The exclusion of the variance in our criterion is also supported by the variance stabilization theory (Durbin, et al., 2002; Huber, et al., 2002), because the geometric mean in (69) is equivalent to the arithmetic mean after logarithmic transformation and the gene expression after logarithmic transformation approximately has equal variance across different genes, especially for the up-regulated genes.

Corresponding to the definition of OVEPUGs, the OVEPDGs (which are down-regulated in one class while being up-regulated in all other classes) can be defined by the following criterion:

$$\mathbb{J}_{\mathrm{PDG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\mathrm{PDG}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \frac{\min_{l \neq k} \{\mu_j(l)\}}{\mu_j(k)} \geq \tau_k \right. \right\}. \tag{70}$$

Furthermore, the combination of PUGs and PDGs can be defined as:

$$\mathbb{J}_{\mathrm{PUG+PDG}} = \bigcup_{k=1}^{M} \mathbb{J}_{\mathrm{PUG+PDG}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \max \left\{ \frac{\mu_j(k)}{\max_{l \neq k} \{\mu_j(l)\}}, \frac{\min_{l \neq k} \{\mu_j(l)\}}{\mu_j(k)} \right\} \geq \tau_k \right. \right\}. \tag{71}$$

Purely from the machine learning view, PDGs have the theoretical capability of being as discriminating as PUGs. Thus, PDGs merit consideration as candidate genes. However, there are several critical differences, with consequential implications, between lowly-expressed genes and highly-expressed genes, such as the extraordinarily large proportion and relatively large noise of the lowly-expressed genes. We have evaluated the classification performance of PUGs, PDGs, and PUGs+PDGs, respectively. Experimental results show that PDGs have less discriminatory

power than PUGs and the inclusion of PDGs actually worsens classification accuracy, compared to just using PUGs. Experiments and further discussion will be given in the results section.

III.1.4 Review of Relevant Gene Selection Methods

Here we briefly review four benchmark gene selection methods that have been previously proposed for multicategory classification, namely, OVRSNR (Golub, et al., 1999), OVR t-statistic (OVRt-stat) (Liu, et al., 2002), BW (Dudoit, et al., 2002), and SVMRFE (Guyon, et al., 2002).

Let $\mu_{j,k}$ and $\mu_{j,-k}$ be the arithmetic means of the expression levels of gene j associated with phenotype k and associated with the super class of remaining phenotypes, respectively, on the log-transformed scale, with $\sigma_{j,k}$ and $\sigma_{j,-k}$ the corresponding standard deviations. OVRSNR gene selection for multicategory classification is given by:

$$\mathbb{J}_{\text{OVRSNR}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{OVRSNR}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \frac{\left|\mu_{j,k} - \mu_{j,-k}\right|}{\sigma_{j,k} + \sigma_{j,-k}} \geq \tau \right. \right\}, \tag{72}$$

where τ is a pre-defined threshold (Golub, et al., 1999). To assess the statistical significance of the difference between $\mu_{j,k}$ and $\mu_{j,-k}$, OVRt-stat applies a test of the null hypothesis that the means of two assumed normally distributed measurements are equal. Accordingly, OVRt-stat gene selection is given by (Liu, et al., 2002):

$$\mathbb{J}_{\text{OVR}t\text{-stat}} = \bigcup_{k=1}^{M} \mathbb{J}_{\text{OVR}t\text{-stat}}(k) = \bigcup_{k=1}^{M} \left\{ j \left| \frac{\left|\mu_{j,k} - \mu_{j,-k}\right|}{\sqrt{\sigma_{j,k}^2 / N_k + \sigma_{j,-k}^2 / \left(N - N_k\right)}} \geq \tau \right. \right\}, \tag{73}$$

where the p-values associated with each gene may be estimated. As aforementioned, one limitation of the gene selection schemes (70) and (71) is that the criterion function considers the remaining classes as a single group. Another is that they both require variance estimation.

Dudiot et al. (Dudoit, et al., 2002) proposed a pooled OVO gene selection method based on the BW sum of squares across all paired classes. Specifically, BW gene selection is specified by

$$\mathbb{J}_{\mathrm{BW}} = \left\{ j \left| \frac{\sum_{i=1}^{N} \sum_{k=1}^{M} \mathbf{1}_{\omega_k}(i)\left(\mu_{j,k} - \mu_j\right)^2}{\sum_{i=1}^{N} \sum_{k=1}^{M} \mathbf{1}_{\omega_k}(i)\left(x_{ij} - \mu_{j,k}\right)^2} \geq \tau \right. \right\}, \tag{74}$$

where μ_j is the global arithmetic center of gene j over all sample points and $\mathbf{1}_{\omega_k}(i)$ is the indicator function reflecting membership of sample i in class k. As pointed out by Loog et al. (Loog, et al., 2001), BW gene selection may only preserve the distances of already well-separated classes rather than neighboring classes (Loog, et al., 2001).

From a dimensionality reduction point of view, Guyon et al. (Guyon, et al., 2002) proposed a feature subset ranking criterion for linear SVMs, dubbed the SVMRFE. Here, one first trains a linear SVM classifier on the full feature space. Features are then ranked based on the magnitude of their weights and are eliminated in the order of increasing weight magnitude. A widely adopted reduction strategy is to eliminate a fixed or decreasing percentage of features corresponding to the bottom portion of the ranked weights and then to retrain the SVM on the reduced feature space. Application to microarray gene expression data shows that the genes selected matter more than the classifiers with which they are paired (Guyon, et al., 2002).

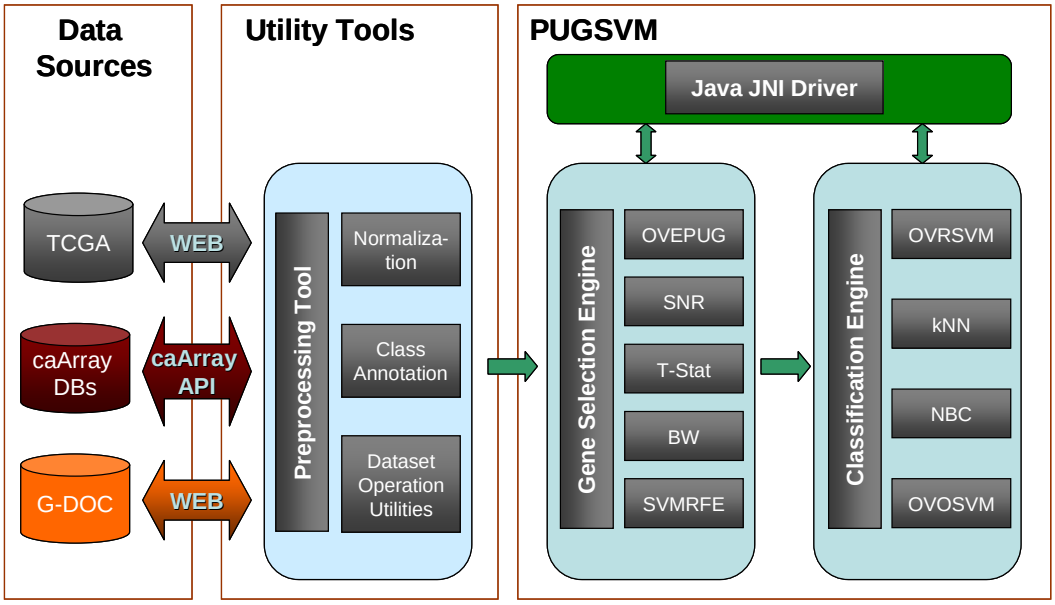

Figure 26: The components and input/output of PUGSVM.

III.2 Software

The components of PUGSVM software and their input/output relationships are illustrated in Figure 26. We use caBIG existing tools to load, preprocess, and normalize gene expression data from in-house (i.e., Georgetown Database of Cancer; GDOC) or public databases (e.g., caArray, TCGA). The processed data with class labeling are fed to the gene selection component. The selected PUGs are then used to train and test the classifiers for predictive classification. The output of PUGSVM is a set of gene markers with good, generalizable performance.

The OVEPUG and other algorithms in the gene selection component are implemented in Matlab. We used the Matlab compiler to generate C++ shared function libraries. The OVRSVM algorithm and other classifiers are implemented in C++ with simple calling interfaces. The user interface is implemented in Java, and C++ shared libraries are called from Java using the Java

Native Interface (JNI). PUGSVM has been tested on Microsoft Windows and Linux platforms. Users can run PUGSVM directly on a computer without an installed version of MATLAB.

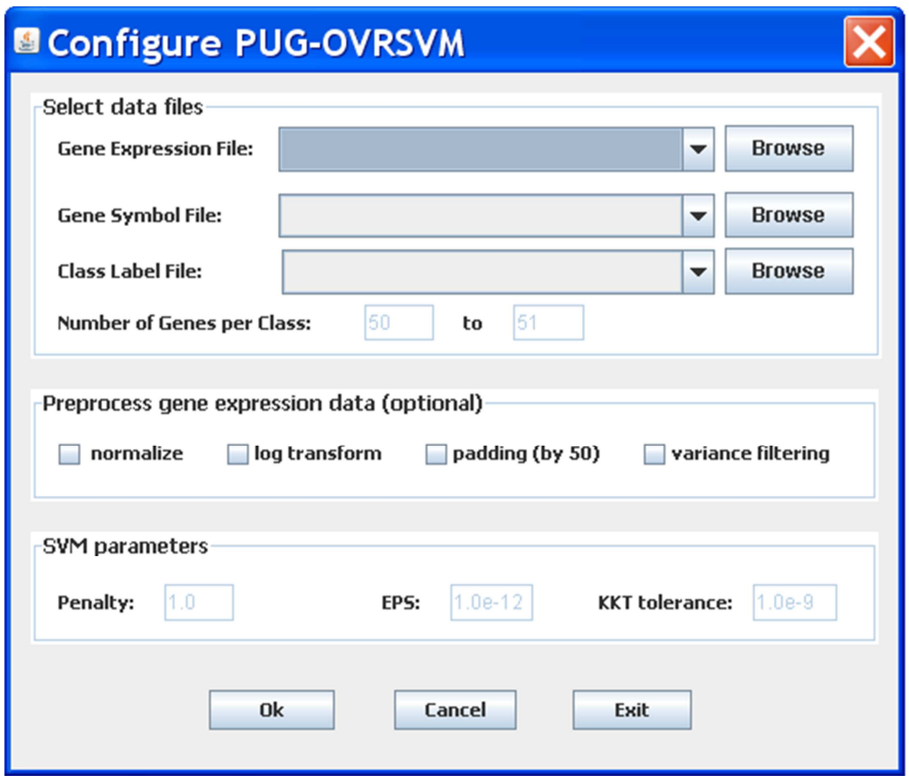

Figure 27: Screenshot of PUGSVM software

Figure 27 shows the screenshot of PUGSVM software. To start the experiment, three files are needed: i) a gene expression file that contains the gene expression value for each probeset either log-transformed or on its original scale; ii) a gene symbol file that provides the name corresponding to each gene in the gene expression data; iii) a class label file that has a column vector of class labels matching each sample in the gene expression data. We also provide the flexibility for users to set their own preprocessing procedure and to adjust the SVM parameters. The output files present (1) the number of correctly classified samples, (2) incorrectly classified samples with their IDs, SVM class (incorrect one), and original class, and (3) a matrix of PUGs

(phenotype up-regulated genes) associated with each class and the corresponding fold-change value in the correctly classified results.

III.3 Experimental Results

We tested PUG-OVRSVM on five benchmark and one in-house real microarray dataset, and compared the performance to several widely-adopted gene selection and classification methods.

III.3.1 Description of Real Datasets

The six datasets are the MIT 14 Global Cancer Map dataset (GCM) (Ramaswamy, et al., 2001), the NCI 60 cancer cell lines dataset (NCI60) (Staunton, et al., 2001), the University of Michigan cancer dataset (UMich) (Shedden, et al., 2003), the Central Nervous System tumors dataset (CNS) (Pomeroy, et al., 2002), the Muscular Dystrophy dataset (MD) (Bakay, et al., 2006), and the Norway Ascites dataset (NAS). To assure a meaningful and well-grounded comparison, we emphasized data quality and suitability in choosing these test datasets. For example, the datasets cannot be too "simple" (if the classes are well-separated, all methods perform equally well) or too "complex" (no method will then perform reasonably well), and each class should contain sufficient samples to support some form of cross-validation assessment.

We also performed several important pre-processing steps widely adopted by other researchers (Guyon, et al., 2002; Ramaswamy, et al., 2001; Shedden, et al., 2003; Statnikov, et al., 2005). When the expression levels in the raw data take negative values, probably due to global probe-set calls and/or data normalization procedures, these negative values are replaced by a fixed small quantity (Shedden, et al., 2003). On the log-transformed scale, we further conducted a variance-

based unsupervised gene filtering operation to remove the genes whose expression standard deviations (across all samples) were less than a pre-determined small threshold; this effectively reduces the number of genes by half (Guyon, et al., 2002; Shedden, et al., 2003).

III.3.2 Experiment Design

We decoupled the two key steps of multicategory classification: 1) selecting an informative subset of marker genes and then 2) finding an accurate decision function. For the crucial first step we implemented five gene selection methods, including OVEPUG specified by (69), OVRSNR specified by (72), OVRt-stat specified by (73), pooled BW specified by (74), and SVMRFE described in (Ramaswamy, et al., 2001). We applied these methods to the six datasets, and for each dataset, we selected a sequence of gene subsets with varying sizes, indexed by N_{subPUG}, the number of genes per class. In our experiments, this number was increased from 2 up to 100. There are several reasons why we do not go beyond 100 subPUGs per class. First, classification accuracy may be either flat or monotonically decreasing as the number of features increases beyond a certain point, due to the theoretical bias-variance dilemma. Second, even in some cases where best performance is achieved using all the gene features, the idea of feature selection is to find the minimum number of features needed to achieve good (near-optimal) classification accuracy. Third, when N_{subPUG} =100, the total number of genes used for classification is already quite large (this number is maximized if the sets $\mathbb{J}_{\text{PUG}}(k)$ are mutually exclusive, in which case it is N_{subPUG} times the number of classes). Fourth, but not least important, a large feature reduction may be necessary not only complexity-wise, but also for interpreting the biological functions and pathway involvement when the selected PUGs are most relevant and statistically reproducible.

The quality of the marker gene subsets was assessed by prediction performance on four subsequently trained classifiers, including OVRSVM, kNN, NBC, and OVOSVM. In relation to the proposed PUG-OVRSVM approach, we evaluated all combinations of these four different gene selection methods and three different classifiers on all six benchmark microarray gene expression datasets.

To properly estimate the accuracy of predictive classification, a validation procedure must be carefully designed, recognizing limits on the accuracy of estimated performance, in particular for small sample size. Clearly, classification accuracy must be assessed on labeled samples 'unseen' during training. However, for multicategory classification based on small, class-imbalanced datasets, a single batch held-out test procedure is not viable, as there will be insufficient samples for both accurate classifier training and accurate validation (Hastie, et al., 2001). A practical alternative is a sound cross-validation procedure, wherein all the data are used for both training and testing, but with held-out samples in a testing fold not used for any phase of classifier training, including gene selection and classifier design (Wang, et al., 2008). In our experiments, we chose LOOCV, wherein a test fold consists of a single sample; the rest of the samples are placed in the training set. Using only the training set, the informative genes are selected and the weights of the linear OVRSVM are fit to the data (Liu, et al., 2005; Shedden, et al., 2003; Yeang, et al., 2001). It is worth noting that LOOCV is approximately unbiased, lessening the likelihood of misestimating the prediction error due to small sample size; however, LOOCV estimates do have considerable variance (Braga-Neto and Dougherty, 2004; Hastie, et al., 2001). We evaluated both the lowest "sustainable" prediction error rate and the lowest prediction error rate,

where the sequence of sustainable prediction error rates were determined based on a moving-average of error rates along the survey axis of the number of genes used for each class, N_{subPUG}, with a moving window of width 5. We also report the number of genes per class at which the best sustainable performance was obtained.

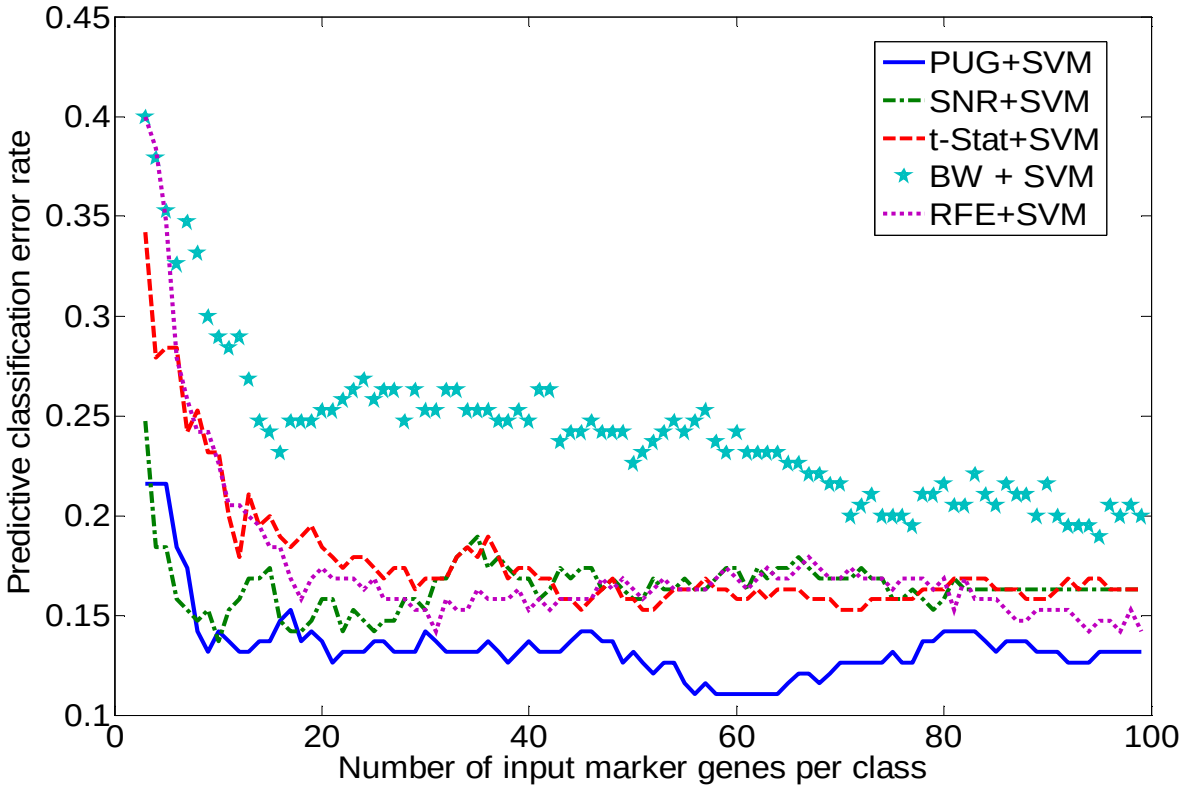

Figure 28: Comparative study on five gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE) using the GCM benchmark dataset. The curves of classification error rates were generated by using OVRSVM committee classifiers with varying size of the input gene subset.

While the error rate is estimated through LOOCV and the optimum number of PUGs used per class is obtained by the aforementioned surveying strategy, we should point out that a two-level LOOCV could be applied to jointly determine the optimum N_{subPUG} and estimate the associated

error rate; however, such an approach is computationally expensive (Statnikov, et al., 2005). For the settings of structural parameters in the classifiers, we used $C = 1.0$ in the SVMs for all experiments (Vapnik, 1998), and chose $k = 1, 3, 5$ in kNNs under different training sample sizes per class, as recommended by (Duda, et al., 2001).

III.3.3 Experimental Results on Real Datasets

Our first comparative study focused on the GCM data widely used for evaluating multicategory classification algorithms (Cai, et al., 2007; Ramaswamy, et al., 2001; Shedden, et al., 2003; Zhou and Tuck, 2007). The performance curves of OVRSVM committee classifiers trained on the commonly pre-processed GCM data using the five different gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE) are detailed in Figure 28. It can be seen that our proposed OVEPUG selection significantly improved the overall multicategory classification when using different numbers of marker genes, as compared to the results produced by the four competing gene selection methods. For example, using as few as 9 genes per phenotypic class (with 126 distinct genes in total, i.e. mutually exclusive PUGs for each class), we classified 164 of 190 (86.32%) of the tumors correctly. Furthermore, using LOOCV on the GCM dataset of 190 primary malignant tumors, and using the optimal number of genes (61 genes per phenotypic class or 769 unique genes in total), we achieved the best (88.95% or 169 of 190 tumors) sustainable correct predictions. In contrast, at its optimum performance, OVRSNR gene selection achieved 85.37% sustainable correct predictions using 25 genes per phenotypic class, OVRt-stat gene selection achieved 84.53% sustainable correct predictions using 71 genes per phenotypic class, BW gene selection achieved 80.53% sustainable correct predictions using 94 genes per phenotypic class, and SVMRFE gene selection achieved 84.74% sustainable correct predictions

using 96 genes per phenotypic class. In our comparative study, instead of solely comparing the lowest error rates achieved by different gene selection methods, we also emphasized the sustainable correct prediction rates, as potential overfitting to the data may produce an (unsustainably) good prediction performance. For our experiments in Figure 28, based on the realistic assumption that the probability of good predictions purely "by chance" over a sequence of consecutive gene numbers is low, we defined the sustainable prediction/error rates based on the moving-averaged prediction/error rates over $\delta = 5$ consecutive gene numbers. Here, δ gives the sustainability requirement.

Table 39: Summary of comparative performances by OVEPUG-OVRSVM and eight competing methods (based on publicly reported optimum results) on the GCM benchmark dataset

References	Gene-select	Classifier	Sample	CV scheme	Error rate
Ramaswamy, *et al.*, 2001	OVRSVM RFE	OVRSVM	144 & 198	LOOCV & 144/54	22.22%
Yeang, 2001	N/A	OVRSVM	144	LOOCV	18.75%
Ooi & Tan, 2003	Genetic algorithm	MLHD	198	144/54	18.00%
Shedden, *et al.*, 2003	OVR fold-change	*k*NN Tree	190	LOOCV	17.37%
Liu, *et al.* 2005	Genetic algorithm	OVOSVM	N/A	LOOCV	20.01%
Statnikov, *et al.*, 2005	No gene selection	CS-SVM	308	10-fold	23.40%
Zhou, *et al.*, 2007	CS-SVM RFE	OVRSVM	198	4-fold	16.72%
Cai, *et al.*, 2007	DISC-GS	*k*NN	190	144/46	21.74%
PUG-OVRSVM	PUG	OVRSVM	190	LOOCV	**11.05%**

For the purpose of information sharing with readers, based on publicly reported optimal results for different methods, we have summarized in Table 39 the comparative performance achieved by

PUG-OVRSVM and eight existing/competing methods on the benchmark GCM dataset, along with the gene selection methods used, the chosen classifiers, sample sizes, and the chosen cross-validation schemes. Obviously, since the reported prediction error rates were generated by different algorithms and under different conditions, any conclusions based on simple/direct comparisons of the reported results must be carefully drawn. We have chosen not to independently reproduce results by re-implementing the methods listed in Table 39, firstly because we typically do not have access to public domain code implementing other authors' methods and secondly because we feel that high reproducibility of previously published results may not be expected without knowing some likely undeclared optimization steps and/or additional control parameters used in the actual computer codes. Nevertheless, many reported prediction error rates on the GCM dataset were actually based on the same/similar training sample set (144~190 primary tumors) and the LOOCV scheme used in our PUG-OVRSVM experiments; furthermore, it was reported that the prediction error rates estimated by LOOCV and 144/54 split/held-out test were very similar (Ramaswamy, et al., 2001). Specifically, the initial work on GCM by Ramaswamy et al. reported an achieved 77.78% prediction rate (Ramaswamy, et al., 2001), and some improved performance was later reported by Yeang et al. (Yeang, et al., 2001) and Liu et al. (Liu, et al., 2002), achieving 81.75% and 79.99% prediction rates, respectively. In the work most closely related to our gene selection scheme, by Shedden et al. (Shedden, et al., 2003), using a kNN tree classifier and using OVR fold-change based gene selection specified by (7), a prediction rate of 82.63% was achieved. In relation to these reported results on GCM, as indicated in Table 39, our proposed PUG-OVRSVM method produced the best sustainable prediction rate of 88.95%, within the LOOCV testing framework.

Table 40: Performance comparison between five different gene selection methods tested on six benchmark microarray gene expression datasets, where the predictive classification error rates for all methods were generated based on OVRSVM committee classification and a LOOCV scheme. Both sustainable and lowest (within parentheses) error rates are reported together with number of genes used per class.

Gene-select	GCM	NCI60	UMich	CNS	MD	NAS
OVE PUG	**11.05%** (11.05%) [61 g/class]	**27.33%** (26.67%) [52 g/class]	**1.08%** (0.85%) [26 g/class]	**7.14%** (7.14%) [71 g/class]	**19.67%** (19.01%) [46 g/class]	**13.16%** (13.16%) [42 g/class]
OVR SNR	14.63% (13.68%) [25 g/class]	31.67% (31.67%) [58 g/class]	1.42% (1.42%) [62 g/class]	**7.14%** (7.14%) [57 g/class]	23.97% (23.97%) [85 g/class]	16.32% (15.79%) [54 g/class]
OVR *t*-stat	15.47% (15.26%) [71 g/class]	31.67% (31.67%) [56 g/class]	1.70% (1.70%) [45 g/class]	7.62% (7.14%) [92 g/class]	23.47% (22.31%) [56 g/class]	15.79% (15.79%) [74 g/class]
BW	19.47% (18.95%) [94 g/class]	31.67% (31.67%) [55 g/class]	1.30% (1.13%) [92 g/class]	**7.14%** (7.14%) [56 g/class]	19.83% (19.01%) [71 g/class]	21.05% (21.05%) [65 g/class]
SVM RFE	15.26% (14.21%) [96 g/class]	29.00% (28.33%) [81 g/class]	1.13% (1.13%) [58 g/class]	14.29% (14.29%) [53 g/class]	29.09% (28.10%) [73 g/class]	32.11% (31.58%) [94 g/class]

A more stringent evaluation of the robustness of a classification method is to carry out the predictions on multiple datasets and then assess the overall performance (Statnikov, et al., 2005). Our second comparative study evaluated the aforementioned five gene selection methods using the six benchmark microarray gene expression datasets. To determine whether the genes selected matter more than the classifiers used (Guyon, et al., 2002), we used a common OVRSVM committee classifier and LOOCV scheme in all the experiments, and summarized the corresponding results in Table 40. For each experiment that used a distinct gene selection

scheme applied to a distinct dataset, we reported both sustainable (with sustainability requirement $\delta = 5$) and lowest (within parentheses) prediction error rates, as well as the number of genes per class that were used to produce these results. Clearly, the selected PUGs based on (8) produced the highest overall sustainable prediction rates as compared to the other four competing gene selection methods. Specifically, PUG is the consistent winner in 22 of 24 competing experiments (combinations of four gene selection schemes and six testing datasets). It should be noted that although BW and OVRSNR achieved comparably low prediction error rates on the CNS dataset (with relatively balanced mixture distributions), they also produced high prediction error rates on the other testing datasets; the other competing gene selection methods also show some level of performance instability across data sets.

Table 41: Performance comparison based on the lowest predictive classification error rates produced by OVEPUG-OVRSVM and the optimum combinations of five different gene selection methods and three different classifiers, tested on six benchmark microarray gene expression datasets and assessed via the LOOCV scheme.

	GCM	NCI60	UMich	CNS	MD	NAS
OVR SVM	**11.05%** (OVEPUG)	**27.33%** (OVEPUG)	**1.08%** (OVEPUG)	**7.14%** (OVEPUG)	**19.67%** (OVEPUG)	**13.16%** (OVEPUG)
OVO SVM	14.74% (OVEPUG)	33.33% (OVRSNR)	1.70% (OVEPUG)	9.52% (BW)	19.83% (BW)	16.32% (OVRSNR)
*K*NN	21.05% (OVEPUG)	31.67% (OVR*t*-stat)	2.27% (OVEPUG)	13.33% (OVEPUG)	21.81% (BW)	13.68% (OVR*t*-stat)
NBC	36.00% (OVRSNR)	51.67% (OVRSNR)	2.83% (OVR*t*-stat)	37.62% (BW)	37.69% (BW)	34.21% (OVEPUG)

To give more complete comparisons that also involved different classifiers (Statnikov, et al., 2005), we further illustrate the superior prediction performance of the matched OVEPUG selection and OVRSVM classifier as compared to the best results produced by combinations of three different classifiers (OVOSVM, kNN, NBC) and four gene selection methods (PUG, OVRSNR, OVRt-stat, pooled BW). The optimum experimental results achieved over all combinations of these methods on the six datasets are summarized in Table 41, where we report both sustainable prediction error rates and the corresponding gene selection methods. Again, PUG-OVRSVM outperformed all other methods on all six datasets and was a clear winner in all 15 competing experiments. Our comparative studies also reveal that although gene selection is a critical step of multi-category classification, the classifiers used do indeed play an important role in achieving good prediction performance.

III.3.4 Comparison Results on Realistic Simulation Datasets

To more reliably validate and compare the performance of the different gene selection methods, we have conducted additional experiments involving realistic simulations. The advantage of using synthetic data is that, unlike the real datasets, often with small sample size and with LOO as the only applicable validation method, large testing samples can be generated to allow an accurate and reliable assessment of a classifier's generalization performance. Two different simulation approaches were implemented. In both, we model the joint distribution for microarray data under each class and generate *i.i.d.* synthetic datasets consistent both with these distributions and with assumed class priors. In the first approach, we chose the class-conditional models consistent with commonly accepted properties of microarray data (few discriminating features, many non-discriminating features, and with small sample size) (Hanczar and Dougherty, 2010;

Wang, et al., 2002). In the second approach, we directly estimated the class-conditional models based on a real microarray dataset and then generated *i.i.d.* samples according to the learned models.

III.3.4.A Design I

We simulated 5000 genes, with 90 "relevant" and 4910 "irrelevant" genes. Inspired by the gene clustering concept in modeling local correlations, we divided the genes into 1000 blocks of size five, each containing exclusively either relevant or irrelevant genes. Within each block the correlation coefficient is 0.9, with zero correlation across blocks. Irrelevant genes are assumed to follow a (univariate) standard normal distribution, for all classes. Relevant genes also follow a normal distribution with variance 1 for all classes. There are three equally likely classes, A, B and C. The mean vectors of the 90 relevant genes under each class are shown in Table 42. The means were chosen to make the classification task neither too easy nor too difficult and to simulate unequal distances between the classes -- A and B are relatively close, with C more distant from both A and B.

Table 42: The mean vectors of the 90 relevant genes under each of the three classes

	The mean vector μ for each class
μ_A	[2.8 2.8 2.8 2.8 2.8 1 1 1 1 1 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 1 1 1 1 1 3 3 3 3 3 0.1 0.1 0.1 0.1 0.1]
μ_B	[1 1 1 1 1 2.8 2.8 2.8 2.8 2.8 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0.5 0.5 0.5 0.5 0.5 0 0 0 0 0 1 1 1 1 1 3 3 3 3 3 0.1 0.1 0.1 0.1 0.1]
μ_C	[1 1 1 1 1 1 1 1 1 1 14.4 14.4 14.4 14.4 14.4 8.5 8.5 8.5 8.5 8.5 8 8 8 8 8 10 10 10 10 10 10 10 10 10 10 10 9 9 9 9 9 10 10 10 10 10 3 3 3 3 3 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 8.5 8.5 8.5 8.5 8.5 8 8 8 8 8 9 9 9 9 9 11 11 11 11 11 7.1 7.1 7.1 7.1 7.1]

We randomly generated 100 synthetic datasets, each partitioned into a small training set of 60 samples (20 per class) and a large testing set of 6000 samples.

III.3.4.B Design II

The second approach models each class as a more realistic multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$, with the class's mean vector $\boldsymbol{\mu}$ and covariance matrix Σ directly learned from the real microarray dataset GCM. Estimation of a covariance matrix Σ is certainly a challenging task, specifically due to the very high dimensionality of the gene space (p= 15,927 genes in the GCM data) and only a few dozen samples available for estimating $p(p\text{-}1)/2$ free covariate parameters per class. It is also computationally prohibitive to generate random vectors based on full covariances on a general desktop computer. To address both of these problems, we applied a factor model (McLachlan and Krishnan, 2008), which significantly reduces the number of free parameters to be estimated while capturing the main correlation structure in the data.

In factor analysis, the observed $p \times 1$ vector $\mathbf{t}$ is modeled as

$$\mathbf{t} = \boldsymbol{\mu} + \mathbf{W}\mathbf{x} + \boldsymbol{\varepsilon} \qquad (75)$$

where $\boldsymbol{\mu}$ is the mean vector of observation $\mathbf{t}$, $\mathbf{W}$ is a $p \times q$ matrix of factor loadings, $\mathbf{x}$ is the $q \times 1$ latent variable vector with standard normal distribution $N(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\varepsilon}$ is noise with independent multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Psi})$, $\boldsymbol{\Psi} = \mathbf{diag}(\sigma_1^2, ..., \sigma_p^2)$. The resulting covariance matrix Σ is

$$\Sigma = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}. \qquad (76)$$

Estimation of Σ reduces to estimating $\mathbf{W}$ and $\mathbf{\Psi}$, totaling $p(q+1)$ parameters. Usually, we have q much less than p. The factor model is learned via the EM algorithm (McLachlan and Krishnan, 2008), initialized by probabilistic principal component analysis (Tipping and Bishop, 1999).

In our experiments, we set $q = 5$, which typically accounted for 60% of the energy. We also tried q=3 and 7 and observed that the relative performance remained unchanged, although the absolute performance of all methods does change with q.

Five phenotypic classes were used in our simulation: breast cancer, lymphoma, bladder cancer, leukemia and CNS. 100 synthetic datasets were generated randomly according to the learned class models from the real data of these five cancer types. The dimension for each sample is 15,927. For each dataset, the training sample size was the same as used in the real data experiments, with 11, 22, 11, 30, and 20 samples in the five respective classes; and the testing set consisted of 3,000 samples, 600 per class.

III.3.4.C Evaluation of performance

For a given gene-selection method and for each data set (indexed by i=1,...,100), the classifier F_i is learned. We then evaluate F_i on the i-th testing set, and measure the error rate ε_i. Since the testing set has quite large sample size, we would expect ε_i to be close to the true classification error rate for F_i. Over 100 simulation datasets, we then calculated both the average classification error $\bar{\varepsilon}$ and the standard deviation δ.

Furthermore, let $\varepsilon_{i,PUG}$ denote the error rate associated with PUGs on testing set *i*, and similarly, let $\varepsilon_{i,SNR}$, ε_{i,t_stat}, $\varepsilon_{i,BW}$ and $\varepsilon_{i,SVMRFE}$ denote the error rates associated with the five peer gene selection methods. The error rate difference between two methods, *e.g.* PUG and SNR, is defined by

$$D_i\left(PUG, SNR\right) = \varepsilon_{i,PUG} - \varepsilon_{i,SNR} \tag{77}$$

For each synthetic dataset, we define the "winner" as the one with least testing error rate. For each method, the mean and standard deviation of the error rate and the frequency of winning are examined for performance evaluation. In addition, the histogram of error rate differences between PUG and peer methods are provided.

III.3.4.D Experimental Results on Simulation Datasets

We tested all gene selection methods using the common OVRSVM classifier. All the experiments were done using the same procedure as on the real datasets, except with LOOCV error estimation replaced by the error estimation using large size independent testing data. Figure 29, analogous to Figure 28 while on the realistic synthetic data whose model was estimated from GCM dataset (simulation data under design II), shows the comparative study on five gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE). Table 43 and Table 44 show the average error, standard deviation, and frequency of winning, estimated based on the 100 simulation datasets. PUG has the smallest average error over all competing methods. PUG also achieves the most stable method (with smallest standard deviation). Table 45 and Table 46 provide the comparison results of the five competing methods on the first ten datasets.

Figure 30 and Figure 31 show histograms of the error difference between PUG and other methods, where a negative value of the difference indicates better performance by PUG. The red bar shows the position where the two methods are equal. We can see that the vast majority of differences are negative. Actually, as indicated in Table 43 and Table 44, there is no positive difference in the subfigures of Figure 30 and at most one positive difference in the subfigures of Figure 31.

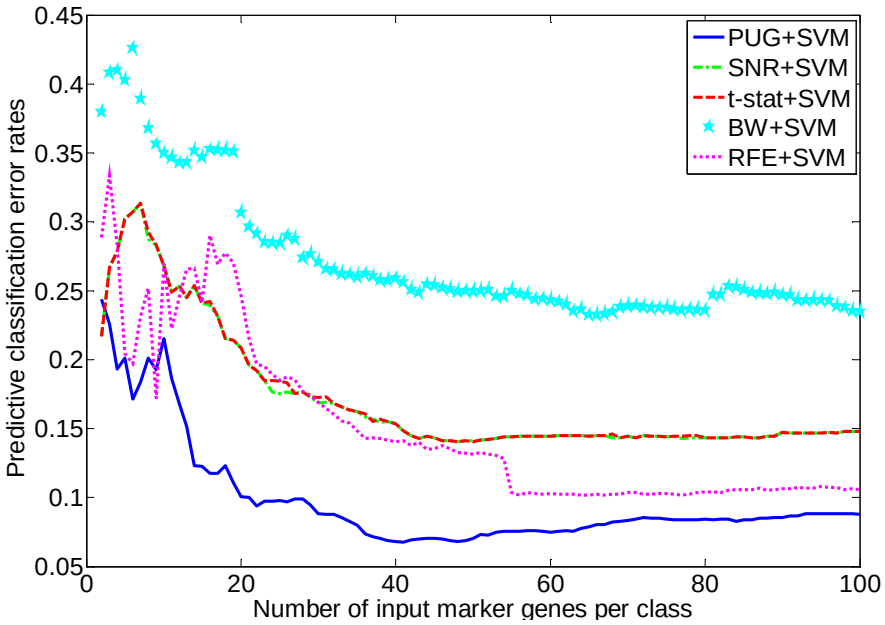

Figure 29: Comparative study on five gene selection methods (OVEPUG, OVRSNR, OVRt-stat, BW, and SVMRFE) on one simulation dataset under design II. The curves of classification error rates were generated by using OVRSVM committee classifiers with varying size of the input gene subset.

Table 43: The mean and standard deviation of classification error and the frequency of winner based on 100 simulation data sets with design I

	PUG	SNR	t-stat	BW	SVMRFE
mean	**0.0724**	0.1129	0.1135	0.1165	0.1203
std deviation	**0.0052**	0.0180	0.0188	0.0177	0.0224

frequency of `winner	**100**	0	0	0	0

Table 44: The mean and standard deviation of classification error and the frequency of winner based on 100 simulation data sets with design II

	PUG	SNR	t-stat	BW	SVMRFE
mean	**0.0712**	0.1311	0.1316	0.2649	0.0910
std deviation	**0.0201**	0.0447	0.0449	0.0302	0.0244
frequency of winner	**99**	0	0	0	1

Table 45: Comparison of the classification error for the first ten simulation datasets with design I

	sim_1	sim_2	sim_3	sim_4	sim_5	sim_6	sim_7	sim_8	sim_9	sim_10
PUG	**0.0864**	**0.0773**	**0.0697**	**0.0681**	**0.0740**	**0.0761**	**0.0740**	**0.0721**	**0.0666**	**0.0758**
SNR	0.1078	0.1092	0.1028	0.1279	0.1331	0.1004	0.1011	0.1253	0.0817	0.0838
t-stat	0.1109	0.1089	0.1022	0.1251	0.1333	0.0991	0.1016	0.1268	0.0823	0.0832
BW	0.1127	0.0995	0.1049	0.1271	0.1309	0.1107	0.1044	0.1291	0.0903	0.0845
SVMRFE	0.1030	0.1009	0.0967	0.1219	0.1248	0.1016	0.1107	0.1191	0.1198	0.0933

Table 46: Comparison of the classification error for the first ten simulation datasets with design II

	sim_1	sim_2	sim_3	sim_4	sim_5	sim_6	sim_7	sim_8	sim_9	sim_10
PUG	**0.0694**	**0.0610**	**0.0748**	**0.0675**	0.0536	**0.0474**	**0.0726**	**0.0818**	**0.0560**	**0.0700**
SNR	0.1559	0.0659	0.1142	0.1211	0.0508	0.1937	0.1568	0.1464	0.0797	0.0711
t-stat	0.1559	0.0659	0.1142	0.1210	0.0508	0.1939	0.1568	0.1464	0.0797	0.0712
BW	0.2373	0.2698	0.2510	0.2650	0.3123	0.2464	3070	0.2236	0.2800	0.3055
SVMRFE	0.0906	0.0739	0.864	0.0852	**0.0426**	0.0776	0.0863	0.0973	0.0655	0.0730

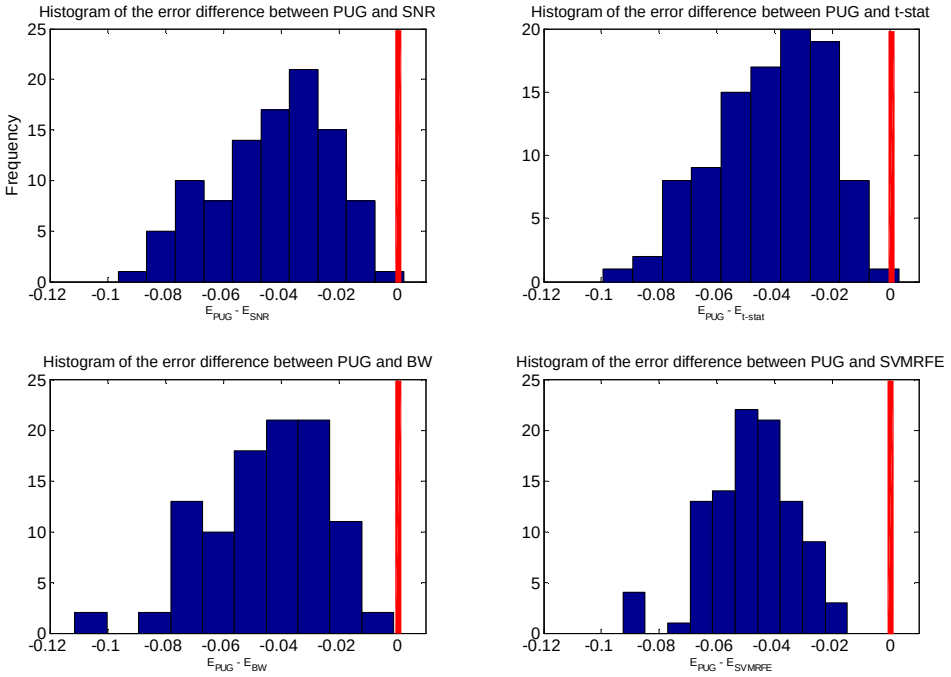

Figure 30: Histogram of the error difference between PUG and other methods with design I.

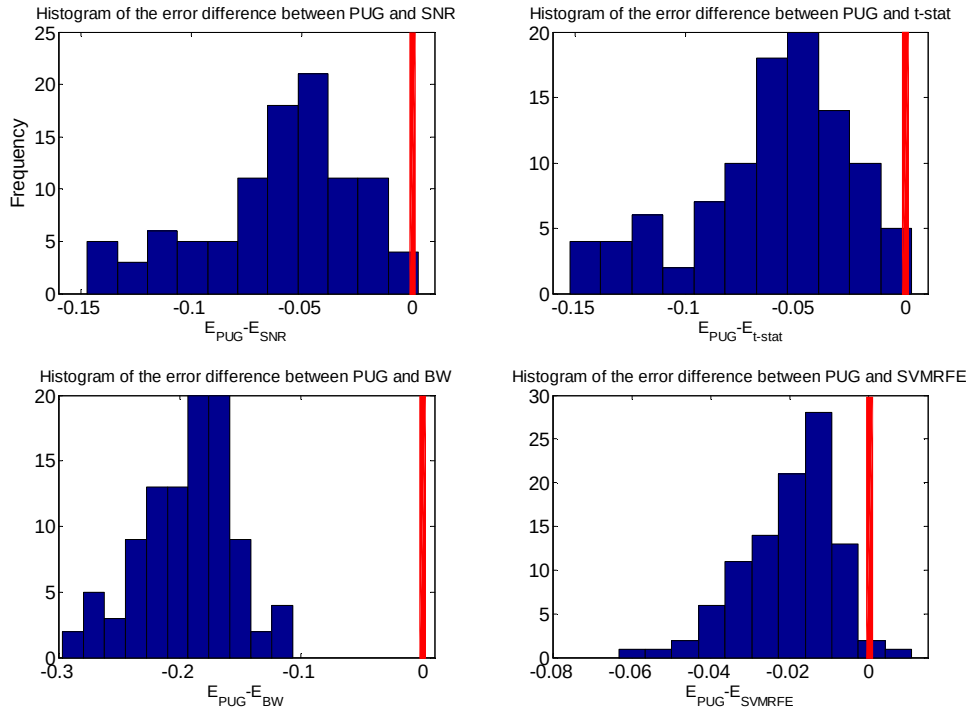

Figure 31: Histogram of the error difference between PUG and other methods with design II.

III.3.4.E Comparison between PUGs and PDGs

In this experiment, we selected PDGs according to the definition given in (9) and evaluated gene selection based on PUGs, PDGs, and based on their union, as given in (10). Again, all gene selection methods were coupled with the OVRSVM classifier. Table 47 shows classification performance for PUGs, PDGs and PUGs+PDGs. Clearly, PDGs have less discriminatory power than PUGs, and the inclusion of PDGs (generally) worsens classification accuracy, compared with just using PUGs.

Table 47: Classification comparison of PUG and PDG on the six benchmark datasets

Error Rate	GCM	NCI60	UMich	CNS	MD	NAS
PUG	**11.05**%	**27.33**%	**1.08**%	**7.14**%	**19.67**%	**13.16**%
PDG	17.58%	30.33%	1.98%	9.52%	26.28%	25.79%
PUG+PDG	14.53%	30.67%	1.13%	**7.14**%	23.14%	15.79%

There are several potential reasons that may jointly explain the non-contributing or even negative role of the included PDGs. First, the number of PDGs are much less than that of PUGs, i.e., PUGs represent the significant majority of informative genes when PUGs and PDGs are jointly considered, as shown in Table 48 (Top PUG+PDGs were selected with 10 genes per class and we counted how many PUGs are included in the total). Second, PDGs are less reliable than PUGs due to the noise characteristics of gene expression data, i.e., low gene expressions contain relatively large additive noise after log-transformation (Huber, et al., 2002; Rocke and Durbin, 2001). This is further exacerbated by the follow-up one-versus-rest classifier because there are many more samples in the ‘rest’ group than in the ‘one’ group. This practically increases the relative noise/variability associated with PDGs in the ‘one’ group. In addition, PUGs are

consistent with the practice of molecular pathology and thus may have broader clinical utility, e.g., most currently available disease gene markers are highly expressed (Shedden, et al., 2003).

Table 48: The percentage of PUGs in the PUG+PDG selection on the six benchmark datasets

	GCM	NCI60	UMich	CNS	MD	NAS
No. of PUG	113	76	56	33	76	65
No. of PUG+PDG	140	90	60	50	130	70
% of PUG	80.71%	84.44%	93.33%	66.00%	58.46%	92.86%

III.3.5 Marker Gene Validation by Biological Knowledge

We have applied existing biological knowledge to validate biological plausibility of the selected PUG markers for two datasets, GCM and NAS.

III.3.5.A Biological interpretation for GCM dataset

Prolactin-induced protein, which is regulated by prolactin activation of its receptors, ranks highest among the PUGs associated with breast cancer. Postmenopausal breast cancer risk is strongly associated with elevated prolactin levels (PubMed IDs 15375001, 12373602, 10203283). Interestingly, prolactin release proportionally increases with increasing central fat in obese women (PubMed ID 15356045) and women with this pattern of obesity have an increased risk of breast cancer mortality (PubMed ID 14607804). Other genes of interest that rank among the top 10 breast cancer PUGs include CRABP2, which transports retinoic acid to the nucleus. Retinoids are important regulators of breast cell function and show activity as potential breast cancer chemopreventive agents (PubMed IDs 11250995, 12186376). Mammglobin is primarily expressed in normal breast epithelium and breast cancers (PubMed ID 12793902). Carbonic anhydrase XII is expressed in breast cancers and is generally considered a marker of a good

prognosis (PubMed ID 12671706). The selective expression and/or function of these genes in breast cancers are consistent with their selection as PUGs in the classification scheme.

The top 10 PUGs associated with prostate cancer include several genes strongly associated with the prostate including prostate specific antigen (PSA) and its alternatively spliced form 2, and prostatic secretory protein 57. The role of PSA gene KLK3 and KLK1 as a biomarker of prostate cancer is well established (PubMed ID 19213567). Increased NPY expression is associated with high-grade prostatic intraepithelial neoplasia and poor prognosis in prostate cancers (PubMed ID 10561252). ACPP is another prostate specific protein biomarker (PubMed ID 8244395). The strong representation of genes that show clear selectivity for expression within the prostate illustrates the potential of the PUGs as bio-markers linked to the biology of the underlying tissues.

Several of the selected PUG markers for uterine cancer fit very well with our current biological understanding of this disease. It is well-established that estrogen receptor alpha (ESR1) is expressed or amplified in human uterine cancer (PubMed IDs 18720455, 17911007, 15251938), while the Hox7 gene (MSX1) contributes to uterine function in cow and mouse models, especially at the onset of pregnancy (PubMed IDs 7908629, 14976223, 19007558). Mammaglobin 2 (SCGB2A1) is highly expressed in a specific type of well-differentiated uterine cancer (endometrial cancers) (PubMed ID 18021217), and PAM expression in the rat uterus is known to be regulated by estrogen (PubMed IDs 9618561, 9441675). Other PUGs provide novel insights into uterine cancer that are deserving of further study. Our PUG selection ranks HE4 higher than the well-established CA125 marker, which may suggest HE4 as a promising

alternative for the clinical management of endometrial cancer. One recent study (PubMed ID 18495222) shows that, at 95% specificity, the sensitivity of differentiating between controls and all stages of uterine cancer is 44.9% using HE4 versus 25.2% using CA125 ($p = 0.0001$).

Osteopontin (OPN) is an integrin-binding protein that is involved in tumorigenesis and metastasis. OPN levels in the plasma of patients with ovarian cancer are much higher compared with plasma from healthy individuals (PubMed ID 11926891). OPN can increase the survival of ovarian cancer cells under stress conditions in vitro and can promote the late progression of ovarian cancer in vivo, and the survival-promoting functions of OPN are mediated through Akt activation (PubMed ID 19016748). Matrix metalloproteinase 2 (MMP2) is an enzyme degrading collagen type IV and other components of the basement membrane. MMP-2 is expressed by metastatic ovarian cancer cells and functionally regulates their attachment to peritoneal surfaces (PubMed ID 18340378). MMP2 facilitates the transmigration of human ovarian carcinoma cells across endothelial extracellular matrix (PubMed ID 15609323). Glutathione peroxidase 3 (GPX3) is one of several isoforms of peroxidases that reduce hydroperoxides to the corresponding alcohols by means of glutathione (GSH) (PubMed ID 17081103). GPX3 has been shown to be highly expressed in ovarian clear cell adenocarcinoma. Moreover, GPX3 has been associated with low cisplatin sensitivity (PubMed ID 19020706).

III.3.5.B Biological interpretation for NAS dataset

Several top-ranking gene products identified by our computational method have been well established as tumor-type specific markers and many of them have been used in clinical diagnosis. For example, mucin 16, also known as CA125, is an FDA-approved serum marker to

monitor disease progression and recurrence in ovarian cancer patients (PubMed ID 19042984). Likewise, kallikrein family members including KLK6 and KLK8 are known to be ovarian cancer associated markers which can be detected in body fluids in ovarian cancer patients (PubMed ID 17303231). TITF1 (also known as TTF1) has been reported as a relatively specific marker in lung adenocarcinoma (PubMed ID 17982442) and it has been used to assist differential diagnosis of lung cancer from other types of carcinoma. Fatty acid synthase (FASN) is a well-known gene that is often upregulated in breast cancer (PubMed ID 17631500) and the enzyme is amenable for drug targeting using FASN inhibitors, suggesting that it can be used as a therapeutic target in breast cancer. The above findings indicate the robustness of our computational method in identifying tumor-type specific markers and in classifying different types of neoplastic diseases. Such information could be useful in translational studies (PubMed ID 12874019). Metastatic carcinoma of unknown origin is a relatively common presentation in cancer patients and an accurate diagnosis of the tumor type in the metastatic diseases is important to direct appropriate treatment and predict clinical outcome. The distinctive patterns of gene expression characteristic to various types of cancer may help pathologists and clinicians to better manage their patients.

III.3.5.C Gene comparisons between methods

It may be informative to provide some initial analysis on how the selected genes compare between methods; however, without definitive ground truth on cancer markers, the utility of this information is somewhat limited and should, thus, be treated as anecdotal, rather than conclusive. Specifically, we have now done some assessment of how differentially these gene selection methods rank some known cancer marker genes. The overlap rate is defined as the number of genes commonly selected by two methods over the maximum size of the two selected gene sets.

Let G_1 and G_2 denote the gene sets selected by gene selection methods 1 and 2, respectively, and $|G|$ denote the cardinality (the size) of set G. The overlap rate between G_1 and G_2 is

$$R = \frac{|G_1 \cap G_2|}{\max(|G_1|, |G_2|)}. \quad (78)$$

Table 49 shows the overlap rate between methods on the top 100 genes per class. We can see that the overlap rates between methods are generally low except for the pair of SNR and t-stat. BW genes are quite different from the genes selected by all other methods and have only about 15% overlap rate with PUG and SVMRFE. The relatively high overlap rate between SNR and t-stat may be expected since they use quite similar summary statistics in their selection criteria.

Table 49: The overlapping rate between methods on the top 100 genes per class

Overlapping Rate	PUG	SNR	t-stat	BW	SVMRFE
PUG	1	0.4117	0.3053	0.1450	0.4057
SNR	0.4117	1	0.7439	0.3307	0.3841
t-stat	0.3053	0.7439	1	0.2907	0.3941
BW	0.1450	0.3307	0.2907	1	0.1307
SVMRFE	0.4057	0.3841	0.3941	0.1307	1

We have also examined a total of 16 genes with known associations with 4 tumor types. These 16 genes are well-known markers supported by current biological knowledge. The rank of biomedical importance of these genes produced by each method is summarized in Table 50. When a gene is not listed in the top 100 genes by a wrapper method like SVMRFE, we simply assign the rank as '>100'. Generally but not uniformly across cancer types, these validated

marker genes are highly ranked in the PUGs list as compared to other methods, and thus will be surely selected by the PUG criterion.

Table 50: Detailed comparison between methods on several validated marker genes

Breast Cancer Relevant Genes						Prostate Cancer Relevant Genes					
Gene Symbol	Rank					Gene Symbol	Rank				
	PUG	SNR	t-stat	BW	SVMRFE		PUG	SNR	t-stat	BW	SVMRFE
PIP	**1**	5745	6146	473	>100	KLK3	**4**	5	11	61	15
CRABP2	**4**	5965	6244	498	>100	KLK1	**5**	3	9	76	16
SCGB2A2	**6**	6693	6773	458	14	NPY	**7**	18	22	344	30
CA12	**9**	6586	6647	518	>100	ACPP	**3**	4	8	71	12
Uterine Cancer Relevant Genes						Ovarian Cancer Relevant Genes					
Gene Symbol	Rank					Gene Symbol	Rank				
	PUG	SNR	t-stat	BW	SVMRFE		PUG	SNR	t-stat	BW	SVMRFE
ESR1	**1**	2	16	130	5	OPN	**15**	334	517	371	63
Hox7	**2**	4	52	307	12	MMP2	**42**	2626	3045	481	>100
SCGB2A1	**8**	3	19	190	4	GPX3	**7**	411	812	446	>100
PAM	**10**	83	281	365	71						
HE4	**3**	1	3	99	5						

IV. Summary and Future Work

IV.1 Summary of Contributions

IV.1.1 Analysis of Interaction Effects in GWAS

The interaction is a long-existing concept in biology and serves as both an opportunity and a challenge. The incorporation and identification of interaction effects has the potential to improve

detection power, to retrieve the missing heritability in the current GWAS studies, and to indicate the underlying biological processes. The challenges brought by the introduction of the interaction into the analysis are daunting, not only because of the complex form of interactions in complex and common diseases, but also because of the immense computational demand. So, in the process of designing our algorithms, we have always adhered to the following two principles: 1) aiming to increase the testing power as possible as we can, while 2) reducing computation complexity as possible as possible. As we have mentioned previously, we have proposed two objectives taking advantage of interaction effects: (1) incorporating the interaction effect to improve the detect rate of disease-susceptibility markers and (2) determining the interaction effects among markers with significant marginal effects. We have developed systematic approaches to both these objectives, with a series of challenges investigated and addressed to achieve both of these final goals.

For the first objective, i.e., incorporating the interaction effect to improve marker detection power, we proposed a unique and novel coupled criterion, stressing on both high sensitivity and high specificity simultaneously. Our proposed coupled criterion was then skillfully transformed to a constrained optimization problem. Because accurate assessment of significance is a necessity for one to have good testing power for a hypothesis test approach, we devoted a lot of effort to achieving accurate significance assessment. There are quite a few factors that may have a tendency to make assessment either liberal or conservative. The accurate significance assessment for our proposed test was guaranteed by the following efforts. (1) The criterion evaluates a bank of contingency tables and the hypergeometric distribution was adopted as the basic distribution due to its exactness property, which is resistant to singular cells and

imbalanced sample distribution. (2) We proved the non-infinitesimal characteristic of the hypergeometric distribution and proposed an effective solution to correct the effect. Otherwise, the obtained p-value would be conservative. Note that this phenomenon exists in the contingency table irrespective of whether or not we use the hypergeometric distribution, but the hypergeometric distribution makes the correction feasible. (3) The hypergeometric distribution also makes obvious the discreteness of the resulting p-values. Discreteness induces conservative p-values and is corrected by the mid-p-value approach in our algorithm. (4) In general, the correlation structure among multiple comparisons places a significant hurdle on accurate assessment of the p-values for order statistics. However, this is in fact not always the case. We have investigated the characteristics of the correlation structure's influence on the significance assessment and identified the conditions that offer an easy and simple solution. Fortunately, our task satisfies these conditions. (5) We have observed for the first time the phenomenon that the comparison across different orders of interaction may induce conservative p-values. This property was rigorously proved and the correction for such effect was proposed.

For the second objective i.e., determining the interaction effect among significant marginal effect factors, we proposed a criterion inspired by and linked directly to the biological meaning of interaction. We figured out several important practical properties to be satisfied for a criterion identifying interactions with significant marginal effect. The criterion should be compatible or valid in any combination of the following three scenarios: (1) if some real disease-risk factors are not measured, (2) if the observed markers are tag markers instead of the true disease-causing factors, and (3) if the disease is heterogeneous. We point out that the three scenarios occur *quite* commonly, not rarely. Surprisingly, the current gold-standard approach, logistic regression with

interaction terms, fails in all three of these scenarios. Our proposed alternative to logistic regression is valid under all three of these scenarios, as we established through several theorems and corollaries for our proposed approach. Thus, alternative to logistic regression can be applied in practice reliably. We also prove that our approach is more powerful than conventional logistic regression with interaction terms. The significance assessment is obtained by casting our approach under the framework of likelihood ratio test and applying Wilks' theorem. To apply the theory of a likelihood ratio test, the maximum likelihood estimation is a necessary step. It is straightforward to prove that the log likelihood function of our proposed test is convex. Different from the standard generalized linear model is the linear inequality constraints on the parameter space for our model. However, since the function is convex and the constraints are linear inequalities, the maximum likelihood estimation problem is still a classical convex optimization problem. There are a bunch of sophisticated approaches available for this convex problem. We develop an approach that combines Newton method and interior point method, which helps to accelerate convergence.

As we have said before, the huge number of SNPs in GWAS studies poses tremendous computational challenges and reducing the computation complexity is an imperative. It is worth noting that all discussions described here pertain to the first objective, because the markers with significant marginal effects are typically much fewer than the number of SNPs genotyped and the computation for the second objective is generally not a significant burden. We have made two levels of efforts: firstly, reducing the number of candidates to be fully evaluated through an elaborately designed heuristic search scheme; secondly, squeezing the computation time at each step. Let us first discuss the heuristic search. A heuristic search scheme usually concerns the

tradeoff between the computational complexity and the success rate of catching the targets. Besides reduction in the computation time, users also want to know how likely it is that they have missed the targets for a given parameter setting. Unlike most conventional heuristic searching algorithms which heavily rely on only the significance of the marginal effects and whose controlling parameters have no explicit link to the expected performance of catching targets, our proposed search scheme fully explores the characteristics of GWAS and provides a probabilistic framework to control the lower bound of performance, that is, the expected performance under the worst situation. Specifically, we have created the following new concepts and techniques to address the problem. (1) We created a new concept TGEP (*transferable genotype-effect potential*). TGEP has the nice property that the expectation is invariant across different orders of the SNP subset, which makes it possible to make high order inference from lower order information, which hence defines the heuristic search strategy. Actually, TGEP can be considered as defining an optimal combination of information from both the marginal significance and the minor allele frequency, while the approaches that solely rely on the marginal effect result in sub-optimal performance. (2) We created a new concept, 'worst-case situation', to serve as the base-line for performance evaluation. Generally, the performance of the heuristic search depends on the specific characteristics of the future candidates, which are usually unknown unless the future candidate has been searched. Through the concept of 'worst situation', we are able to provide a lower bound on the performance. (3) We proposed an efficient and theoretically-sound heuristic search algorithm based on the seed growing. To detect the d^{th} order interaction, we sequentially grow the first order seeds, the second order seeds, to the $(d\text{-}1)^{th}$ order seeds. Then, candidates for the d^{th} order interaction are constructed by pairing the $(d\text{-}1)^{th}$ order seeds and the rest of SNPs. Two heuristic schemes working together were employed to guide the

search towards the SNP combinations with the most potential to be real interactions. The first scheme was created for the growing of the seeds and the second scheme was designed for the construction of the final candidates for d^{th} order interactions. (4) There are parameters in our approach to control the tradeoff between the computation complexity and the success rate of catching true targets and they were carefully calibrated to meet the final performance requirement, i.e. the lower bound of success rate for catching the targets.

Besides the careful design of a heuristic search strategy as discussed above, the second level of reducing the computation is to squeeze the computing time at each computational step. There are a bunch of optimization measures that we have applied, listed as follows. (1) The coupled criterion for the first objective has the internal capability to reduce the order of search. When there is more than one interaction module, unlike some approaches that need to evaluate up to the number of all the involved SNPs, our approach only requires searching up to the maximum number of SNPs in the given module. (2) The coupled criterion for the first objective also readily suggests an efficient strategy to take advantage of previous computations and results if we are interested in systematically discovering all possible significant SNP subsets ranging from low-order interaction to high-order interaction, instead of evaluating one specified SNP subset only. (3) Though the hypergeometric distribution has a lot of nice properties, its associated computation is one drawback. We developed a fast approach to approximately evaluate, within the allowable fidelity, the hypergeometric distribution. (4) Because GWAS studies usually involve a lot of subjects, the counting is the most computation-intensive part in the whole approach. We created a hierarchical structure to reduce the repetitious counting. (5) We used hash tables to track the status of any SNP subset to avoid the redundant search, since a SNP

subset is the combination of the SNP indexes, instead of a permutation. (6) When we calculate the CDF of the hyper-geometric distribution, we rely on the computation of the binomial coefficients ("*n* choose *k*"). The binomial coefficients are used repeatedly and a direct table is adopted to reduce duplicate computations.

The final result of all these efforts is the creation of software that is applicable to real GWAS studies involving more than one million SNPs and thousands of subjects. Applying our method to both simulation data sets and real GWAS data sets, we have verified and demonstrated the advantages of our proposed algorithms. The experimental results are briefly summarized as the follows. (1) 1000 replication data sets without any ground-truth SNP were simulated to evaluate the type I error, which is equivalent to evaluating the accuracy of the significance assessment. Each data set contains 2000 subjects with 1000 SNPs. The experiments show that SCA has type I error very close to the nominated value, while all the competing methods are far too conservative. The conservativeness increases with the order of interaction. (2) The experiments on 100 simulation data sets with several different interaction models inserted show that SCA is much more powerful than the other competing approaches, quantified by both the sensitivity at allowable false positive rate (FPR) and the area under the receiver operating characteristics (ROC) curve. On a typical data set with 2000 subjects and 1000 SNPs, SCA has area under the ROC curve more than 0.99, while all the other methods have area less than 0.84. At the 0.05 Bonferroni corrected significance level, SCA has an average power of 71.87%, compared to the best power of 31.53% from the other methods. (3) The heuristic search significantly speeds the computation. Without any loss of power, the heuristic search reduces from more than 0.3 million years (the estimated time required for exhaustive search) to only 45 minutes, an improvement of

more than 3 billion fold. (4) We have applied SCA to the SLEGEN datasets, which consist of two independent data sets, LUPUS and LLAS. LUPUS is a dataset composed 317,501 SNPs with Illumania SNP chip. 707 cases and 2318 controls were genotyped in LUPUS data. LLAS has genotyped 8230 SNPs, those extracted from the LUPUS data, and it contains 1760 cases and 2083 controls. Six representative SNPs from six genes/regions are identified to be significantly associated with the SLE with marginal effects. Each pair out of the six regions, totally 15 pairs, was tested for the interaction effects. Among all 15 pairs, two pairs passed the experimental-wise threshold and are considered to be significant. The first pair (IRF5/TNPO3-rs12537284 and HLA region 1 - rs3131379) has a p-value of 8.05×10^{-6}, which is 1.20×10^{-4} after Bonferroni correction. The second pair (BLK/c80rf12-rs7836059 and HLA region 2 - rs9275572) has a p-value of 5.51×10^{-5}, which is 8.26×10^{-4} after Bonferroni correction.

Summarily, we have made the following original contributions to the analysis of interaction effects in GWAS:

1) We have proposed a systematic approach to model and evaluate interaction effects with the objective of finding novel disease-risk factors. A unique and novel coupled criterion is designed to incorporate interaction effects for maker discovery. The criterion, which is inherently time-efficient, possesses both high sensitivity and high specificity with theoretical and empirical evidence. An accurate significance assessment is achieved by overcoming a series of challenges, including the correlation structure for order statistics, discrete and non-infinitesimal p-values, and the cross-order comparison.
2) We have proposed a robust, more highly powered alternative to logistic regression model for detecting statistical interactions between genetic and/or environmental factors of

disease risk. A direct link is established to the biological interaction. Three desirable properties, *i.e.*, being robust to missing factors, surrogates and disease heterogeneity, are pointed out for an approach to be of practical use. These three properties are rigorously proved to be hold for the proposed approach. Moreover, a theoretical comparison to logistical regression model is made and the proposed approach is proved to have more power.

3) We have devised an efficient and user-friendly heuristic search strategy to generate candidates with most potential to have interaction effects. A new concept TGEP (*transferable genotype-effect potential*) is created, which proves to have the nice property that the expectation is invariant across different orders of the SNP subset, making it possible to make high order inference from lower order information. Based on TGEP, a seed growing scheme is designed to get an efficient heuristic search. The worse-case situation is identified to serve as the base-line for performance evaluation. A probabilistic framework is set up to provide a way of setting the parameters to meet the final performance requirement, *i.e.* the lower bound of success rate for catching the targets.
4) We have applied various computational skills to speed up the execution of the proposed algorithm, such as a fast evaluation of the hypergeometric distribution, combining the Newton method and the interior method for constrained convex optimization, hierarchical structure to reduce repetitious counting, and hash table to reduce redundant search.
5) We have applied the proposed approach to a number of real datasets, confirming well-validated interactions with more convincing evidence (generating smaller p-values and requiring fewer samples) than those obtained through conventional methods, eliminating inconsistent results in the original reports, and observing novel discoveries that are

otherwise undetectable. The real application covers a wide spectrum, from gene-gene interaction, gene-environment interaction, to environment-environment interaction, from genome-wide study to candidate-based study, from cancer, heart disease, to lupus.

IV.1.2 Multi-class Gene Selection

On the problem of multi-class gene selection, we propose matched design of the gene selection mechanism and a committee classifier for multicategory molecular classification using microarray gene expression data. A key feature of our approach is to match a simple *one-versus-everyone* (OVE) gene selection scheme to the OVRSVM committee classifier (Ramaswamy, et al., 2001). We focused on marker genes that are highly expressed in one phenotype relative to each of the remaining phenotypes, namely Phenotypic Up-regulated Genes (PUGs). PUGs are identified using the fold change ratio computed between the specified phenotype mean and each of the remaining phenotype means. Thus, we considered a gene to be a marker for the specified phenotype if the average expression associated with this phenotype is high relative to the average expressions in each of the other phenotypes. To assure evenhanded resources for discriminating both neighboring and well-separated classes, we used a fixed number of PUGs for each phenotypic class and pooled all phenotype-specific PUGs together to form a gene marker subset used by the OVRSVM committee classifier. All PUGs referenced by the committee classifier are individually interpretable as potential markers for phenotypic classes, allowing each gene to inform the classifier in a way that is consistent with its mechanistic role (Shedden, et al., 2003). Since PUGs are the union of subPUGs selected by simple univariate OVE fold change analysis, they are expected to be statistically reproducible (Lai, et al., 2006; Shedden, et al., 2003; Shi, et al., 2008).

We tested PUG-OVRSVM on five publicly available benchmark sets and one in-house microarray gene expression dataset and on two simulation datasets, observing significantly improved performance with lower error rates, fewer marker genes, and higher performance stability, as compared to several widely-adopted gene selection and classification methods. The reference gene selection methods were OVRSNR (Golub, et al., 1999), OVRt-stat (Liu, et al., 2002), pooled BW (Dudoit, et al., 2002), and OVRSVM-RFE (Guyon, et al., 2002), and the reference classifiers were kNN, NBC, and one-versus-one (OVO) SVM. With accuracy estimated by leave-one-out cross-validation (LOOCV) (Hastie, et al., 2001), our experimental results showed that PUG-OVRSVM outperforms all combinations of the above referenced gene selection and classification methods in the two simulation datasets and 5 out of the 6 real microarray gene expression datasets, and produced comparable performance on the one remaining dataset. Specifically, tested on the widely-used benchmark microarray gene expression dataset "multicategory human cancers data" (GCM) (Ramaswamy, et al., 2001; Statnikov, et al., 2005), PUG-OVRSVM produced a lower error rate of 11.05% (88.95% correct classification rate) than the best known benchmark error rate of 16.72% (83.28% correct classification rate) (Cai, et al., 2007; Zhou and Tuck, 2007).

Summarily, we have made the following original contributions to gene selection for multicategory disease prediction:

1) We have developed a matched design of the gene selection method and a committee classifier. A key feature of our approach is to match a simple *one-versus-everyone* (OVE) gene selection scheme to the OVRSVM committee classifier.

2) We have proposed a novel gene selection method, OVEPUG, which utilizes the unique characteristics of the multicategory classification problem and possesses many good properties. OVEPUG extends the simple fold change approach, assuring evenhanded resources for discriminating both neighboring and well-separated classes, and selecting statistically reproducible and biologically plausible genes.
3) We have made the proposed algorithm available as a caBIG™ (cancer Biomedical Informatics Grid) analytical tool. The tool is a Java-based, cross-platform implementation, including also several popular gene selection methods and multi-category classifiers.
4) Extensive experiments on both simulations and real applications clearly demonstrated the advantages of our proposed method with lower error rates, fewer marker genes, and higher performance sustainability, as compared to several widely-adopted gene selection and classification methods.

IV.2 Future Work

Several lines are under investigation to extend the current work. The current version of the algorithm takes the case-control study design and uses the discrete genetic/environmental variables as inputs. The extension to handle quantitative traits and continuous variables will be helpful. A direct approach is to apply clustering as a preprocessing step and discretize both the trait and input variable. Another way is to use a continuous probability model to replace the hypergeometric distribution model adopted in the current framework.

As suggested for sources of missing heritability, rare variants and structural variants may also play an important role. Although rare variants are individually ‘rare’, they may have large effect

on disease and collectively explain fairly a large fraction of phenotypic variability. Structural variants usually involve long range of DNA change and thus may have significant impacts on disease. It would be fruitful to include both rare variants and structural variants in the analysis.

In the current design of algorithm, we assume that the data is homogeneous. However, the data possibly comes from different populations, each with unique genetic background. Also, multiple research centers may implement studies independently. The stratification analysis has the potential to relieve effects of confounding factors and generate more reliable results.

The proposed algorithm is purely data-driven. Although it is usually considered as an advantage, sometime people do want to utilize existing biological knowledge. Biological knowledge can be beneficial to specify the form of the interaction and to guide and potentially restrict the search.

Though the computation is reduced significantly from exhaustive search within our proposed framework, there is still large room to further decrease the running time. We also note that the SCA interaction analysis is combinatorial, which can be utilized to parallelize the algorithm and reduce the computation time.

Appendix

Appendix 1.

Proposition 1: **Assume the random variable v follows the hyper-geometric distribution such that the probability $\Pr(v=t)=\binom{j}{t}\binom{M-j}{r-t}\Big/\binom{M}{r}$, where M, j and r are parameters. Denote $\mathrm{H}(b;M,j,r)$ as the p-value associated with the observed value b. Then, the**

minimum possible p-value under the hyper-geometric distribution is, $\min_b\{\mathrm{H}(b;M,j,r)\}=\binom{\max(j,r)}{\min(j,r)}\Big/\binom{M}{\min(j,r)}$ **and the minimum is achieved at** $b=\min(j,r)$**.**

Proof:

Since $\mathrm{H}(b;M,j,r)$ is the p-value associated with the observation b, we have,

$$\mathrm{H}(b;M,j,r)=\sum_{t=b}^{\min(j,r)}\Pr(v=t)$$

$$\mathrm{H}(b;M,j,r)-\mathrm{H}(b-1;M,j,r)=\Pr(v=b)=\binom{j}{b}\binom{M-j}{r-b}\Big/\binom{M}{r}$$

And since $\Pr(v=b)>0$, we obtain $\mathrm{H}(b;M,j,r)-\mathrm{H}(b-1;M,j,r)>0$. $\mathrm{H}(b;M,j,r)$ is strictly monotonically decreasing function. Because the range of b is $\left[\max(0,j+r-M),\min(j,r)\right]$, the minimum of $\mathrm{H}(b;M,j,r)$ achieves when b takes its maximum value,

$$\min_b\{\mathrm{H}(b;M,j,r)\}=\mathrm{H}(\min(j,r);M,j,r)$$

There are two scenarios, $j>r$ or $j\le r$.

When $j>r$, then $\min(j,r)=r$ and we have,

$$\mathrm{H}(\min(j,r);M,j,r)=\mathrm{H}(r;M,j,r)$$

$$=\frac{\binom{j}{r}\binom{M-j}{0}}{\binom{M}{r}}=\frac{\binom{j}{r}}{\binom{M}{r}}=\frac{\binom{\max(j,r)}{\min(j,r)}}{\binom{M}{\min(j,r)}}$$

When $j\le r$, then $\min(j,r)=j$ and we have,

$$\mathrm{H}(\min(j,r);M,j,r)=\mathrm{H}(j;M,j,r)$$

$$=\frac{\binom{j}{j}\binom{M-j}{r-j}}{\binom{M}{r}}=\frac{\binom{M-j}{r-j}}{\binom{M}{r}}=\frac{\binom{M-j}{M-r}}{\binom{M}{M-r}}$$

$$= \frac{\frac{(M-j)!}{(M-r)!(M-j-M+r)!}}{\frac{M!}{(M-r)!r!}} = \frac{\frac{(M-j)!}{(r-j)!}}{\frac{M!}{r!}}$$

$$= \frac{\frac{r!}{(r-j)!}}{\frac{M!}{(M-j)!}} = \frac{\frac{r!}{j!(r-j)!}}{\frac{M!}{j!(M-j)!}} = \frac{\binom{r}{j}}{\binom{M}{j}} = \frac{\binom{\max(j,r)}{\min(j,r)}}{\binom{M}{\min(j,r)}}$$

Q.E.D. ■

Appendix 2. ***Theorem 1:*** **Assume** $\mathrm{H}(b;M,j,r)$ **is the p-value associated with the observed value** b **under the hyper-geometric distribution of parameters** (M,j,r)**. Denote the early-stop estimation as** $\hat{\mathrm{H}}(b;M,j,r) = \sum_{t=b}^{\min(j,r,t_s)} p(t;M,j,r)$ **for** $b \geq \mu + 2\sigma$ **, where** $\mu = \frac{j \times r}{M}$ **,** $\sigma^2 = \frac{jr(M-j)(M-r)}{M^2(M-1)}$ **and** t_s **is the first** t **such that** $p(t;M,j,r) \leq 0.05p(b;M,j,r)$**. Then,** $0.95\mathrm{H}(b;M,j,r) \leq \hat{\mathrm{H}}(b;M,j,r) \leq \mathrm{H}(b;M,j,r)$ **and** $t_s - b \leq 1.16\sigma$**.**

Proof:

Let z denote the random variable that follows the standard normal distribution. We write the normal probability density function as,

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

and write the probability of taking values above z for some certain value z is,

$$F(z) = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Denote $z_s = \frac{t_s - \mu}{\sigma}$ and $z_b = \frac{b-\mu}{\sigma}$. Then, we have,

$$p(t;M,j,r) \approx f\left(\frac{t-\mu}{\sigma}\right),$$

$$\mathrm{H}(b;M,j,r) \approx F(z_b),$$

and $$\hat{\mathrm{H}}(b;M,j,r) \approx F(z_b) - F(z_s).$$

From the theory of normal distribution, we know that $F(z)$ has the following inequality(Durrett, 1996),

$$\left(1-\frac{1}{z^2}\right)\frac{1}{z} f(z) \le F(z) \le \frac{1}{z} f(z)$$

Thus, for $z \ge 2$, with the relative error at most 1.3%, $F(z)$ can be approximated (from our own research results, not shown here) as,

$$F(z) = g(z) f(z),$$

where $g(z) = \left(1-\frac{0.673}{z^2}\right)\frac{1}{z}$.

The function $g(z)$ is a decreasing function as its derivative is less than 0 for $z \ge 2$,

$$\frac{d}{dz}(g(z)) = -\left(\frac{1}{z^2} - \frac{0.673\times 3}{z^4}\right) = -\frac{1}{z^2}\left(1-\frac{2.019}{z^2}\right) < 0.$$

Since t_s is the first t such that $p(t;M,j,r) \le 0.05p(b;M,j,r)$, we have

$$f(z_s) \approx 0.05 f(z_b).$$

And because $z_s \ge z_b$ and $g(z)$ is decreasing,

$$g(z_s) \le g(z_b).$$

Hence, we can get that,

$$\begin{aligned} & F(z_s) = g(z_s) f(z_s) \\ & \le g(z_b) f(z_s) \approx 0.05 g(z_b) f(z_b) \\ & = 0.05 F(z_b) \end{aligned}$$

Thus,

$$\begin{aligned}
&\hat{\mathrm{H}}(b;M,j,r) \approx F(z_b) - F(z_s) \\
&\le F(z_b) - 0.05F(z_b) \\
&= 0.95F(z_b) \\
&= 0.95\mathrm{H}(b;M,j,r)
\end{aligned}$$

Obviously, $\hat{\mathrm{H}}(b;M,j,r) = \mathrm{H}(b;M,j,r) - \sum_{t=t_s}^{\min(j,r)} p(t;M,j,r) \le \mathrm{H}(b;M,j,r)$.

Since $f(z_s) \approx 0.05 f(z_b)$ and $z_b \ge 2$, we have,

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{z_s^2}{2}} \approx 0.05 \frac{1}{\sqrt{2\pi}} e^{-\frac{z_b^2}{2}}$$

Then, we can obtain that,

$$z_s^2 - z_b^2 \approx -2\ln(0.05) = 5.99 .$$

Denoting $y = z_s - z_b$, then we get,

$$y(y + 2z_b) = 5.99 .$$

Solving the quadratic equation, we have,

$$y = -z_b + \sqrt{z_b^2 + 5.99} \qquad (\because y \ge 0)$$

Since $z_b \ge 2$, the maximum of y is obtained when $z_b = 2$,

$$\max\{y\} = 1.16 .$$

Thus,

$$t_s - b = (z_s - z_b)\sigma \le 1.16\sigma$$

Q.E.D. ■

Appendix 3. *Theorem 2:* Assume $\mathrm{H}(b;M,j,r)$ is the p-value associated with the observed value b under the hyper-geometric distribution of parameters (M,j,r). Define the function

$\Gamma(r) = \mathrm{H}(\min(j,r); M, j, r)$. **Then,** $\Gamma(r)$ **is strictly decreasing for** $r \le j$**, strictly increasing for** $r > j$**, and achieves its minimum at** $r = j$**.**

Proof:

From ***Proposition 1,*** we know that,

$$\Gamma(r) = \mathrm{H}(\min(j,r); M, j, r) = \binom{\max(j,r)}{\min(j,r)} \Big/ \binom{M}{\min(j,r)}$$

Now check the value of $\Gamma(r)/\Gamma(r-1)$.

For $r \le j$, we have $\max(j,r) = j$ and $\min(j,r) = r$, then,

$$\begin{aligned}
&\Gamma(r)/\Gamma(r-1) \\
&= \left(\binom{\max(j,r)}{\min(j,r)} \Big/ \binom{M}{\min(j,r)} \right) \div \left(\binom{\max(j,r-1)}{\min(j,r-1)} \Big/ \binom{M}{\min(j,r-1)} \right) \\
&= \left(\binom{j}{r} \Big/ \binom{M}{r} \right) \div \left(\binom{j}{r-1} \Big/ \binom{M}{r-1} \right) \\
&= \left(\frac{j!}{r!(j-r)!} \Big/ \frac{M!}{(r)!(M-r)!} \right) \div \left(\frac{j!}{(r-1)!(j-r+1)!} \Big/ \frac{M!}{(r-1)!(M-r+1)!} \right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{(j-r+1)!}{(j-r)!} \times \frac{(M-r)!}{(M-r+1)!} \\
&= \frac{(j-r+1)}{(M-r+1)} \\
&< 1 \qquad\qquad (\because j < M)
\end{aligned}$$

So, $\Gamma(r)$ is strictly decreasing for $r \le j$.

Similarly, for $r > j$, we have $\max(j,r) = r$ and $\min(j,r) = j$, then,

$$\Gamma(r)/\Gamma(r-1)$$

$$= \left(\binom{\max(j,r)}{\min(j,r)} \Big/ \binom{M}{\min(j,r)} \right) \div \left(\binom{\max(j,r-1)}{\min(j,r-1)} \Big/ \binom{M}{\min(j,r-1)} \right)$$

$$= \left(\binom{r}{j} \Big/ \binom{M}{j} \right) \div \left(\binom{r-1}{j} \Big/ \binom{M}{j} \right) = \binom{r}{j} \div \binom{r-1}{j}$$

$$= \frac{r!}{j!(r-j)!} \div \frac{(r-1)!}{j!(r-j-1)!}$$

$$= \frac{r}{r-j}$$

$$> 1 \qquad (\because 0 < j < r)$$

So, $\Gamma(r)$ is strictly increasing for $r > j$.

Since $\Gamma(r)$ decreases before $r = j$ and increases after $r = j$, a simple inference is that $\Gamma(r)$ achieves its minimum at $r = j$.

Q.E.D ■

Appendix 4. *Corollary 2.1:* There exists a solver of $\Gamma^{-1}(P_\Phi)$ with logarithm computation complexity.

Proof:

From ***Theorem 2,*** we know $\Gamma(r)$ is a strictly decreasing function for $0 \le r \le j$. Supposing $\Gamma(a) > P_\Phi$ and $\Gamma(b) \ge P_\Phi$ with $a < b$. The monotonicity of $\Gamma(r)$ makes us apply the binary search algorithm, that is, update the pair $\{a, b\}$ iteratively until $b - a = 1$. Denote $t = \lfloor (a+b)/2 \rfloor$, where $\lfloor x \rfloor$ takes the nearest integer equal to or less than x. The rule of update is,

$$\begin{cases} a = t & \textbf{if} \quad \Gamma(t) > P_\Phi \\ b = t & \textbf{if} \quad \Gamma(t) \le P_\Phi \end{cases}$$

From ***Theorem 2,*** we know $\Gamma(r)$ has minimum at $r = j$. Obviously, As a p-value, $\Gamma(r)$ has maximum of 1 at $r = 0$. Then, to initialize the program, we set $a = 0$ and $b = j$.

From the theory of binary search algorithm(Knuth, 1997), we know that the proposed approach operates in $O(\log_2(j))$, that is, in the worst case it will operate in the order of $\log_2(j)$.

Q.E.D ■

Appendix 5. *Theorem 3:* Assume the vector X of *n* random variables with zero means and unit variances follow a multivariate normal distribution with the covariance matrix as *C*, which has elements $\{\rho_{ij}\}$. Denote $Y = \max\{X_1, X_2, ..., X_n\}$ and $\Phi(\bullet)$ the CDF of standard normal distribution. Write $R = \Pr(Y \geq y) / (n(1-\Phi(y)))$. Then, we have, (1), *R* is an increasing function w.r.t. *y*; (2), $\lim_{y\to\infty} R = 1$ if $\rho_{ij} < 1$; and (3) $R \geq \frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - \sum_{j=1, j\neq i}^{n} \Phi\left(-y \times \sqrt{(1-\rho_{ij})/(1+\rho_{ij})}\right)\right)$.

Proof:

Part (1):

The cumulative distribution function of *Y* can be written as,

$$\Pr(Y \leq y) = \Pr(\max\{X_1, X_2, ..., X_n\} < y)$$

$$= \Pr(X_1 < y, X_2 < y, ..., X_n < y)$$

Denote $\phi(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \mathbf{0}, \mathbf{C})$ the probability density function (pdf) of multivariate normal distribution with zero means $\mathbf{0}$ and covariance matrix $\mathbf{C}$. In the following discussions, without confusion, we will write $\phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C})$ in replace of $\phi(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \mathbf{0}, \mathbf{C})$. So, $\Pr(Y \leq y)$ can be expressed as,

$$\Pr(Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{y}...\int_{-\infty}^{y} \phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C})\, dX_n ... dX_2 dX_1$$

Applying the multivariable chain rule and taking the derivative with respect to *y*, we obtain the probability density function for *Y*,

$$f(Y = y) = \frac{d}{dy}\int_{-\infty}^{y}\int_{-\infty}^{y}...\int_{-\infty}^{y} \phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C})\, dX_n ... dX_2 dX_1$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \int_{-\infty}^{y} ... \int_{-\infty}^{y_i} ... \int_{-\infty}^{y} \phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C}) dX_n ... dX_i .. dX_1 \Big|_{y_i = y}$$

Conditioned on the *i*th variable, the multivariate normal pdf $\phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C})$ can be written as,

$$\phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C})$$
$$= \phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C} \mid X_i = x_i) \phi(x_i; \mathbf{0}, \mathbf{1})$$

Rewrite the covariance matrix $\mathbf{C}$,

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{-i,-i}, & \mathbf{C}_{-i,n} \\ \mathbf{C}_{n,-i}, & \mathbf{C}_{i,i} \end{bmatrix}$$

where $\mathbf{C}_{-i,-i}$ is a $(n-1)$-by-$(n-1)$ matrix excluding the *i*th row and the *i*th column of **C,** $\mathbf{C}_{-i,n}$ is a $(n-1)$ dimensional column vector and the *i*th column of **C** excluding the *i*th element, $\mathbf{C}_{n,-i}$ is a $(n-1)$ dimensional row vector and the *i*th row of **C** excluding the *i*th element, $\mathbf{C}_{i,i}$ is the element of C at the *i*th row and *i*th column.

From the theory of multivariate normal distribution, we know that the conditional distribution is also multivariate normal with the conditional mean $\boldsymbol{\mu}_{c,i}$ and the conditional covariance $\mathbf{C}_{c,i}$ on the *i*th variable $X_i = x_i$, where we have the formulas for $\boldsymbol{\mu}_{c,i}$ and $\mathbf{C}_{c,i}$,

$$\boldsymbol{\mu}_{c,i} = \mathbf{C}_{-i,n} x_i$$
$$\mathbf{C}_{c,i} = \mathbf{C}_{-i,-i} - \mathbf{C}_{-i,n} \mathbf{C}_{n,-i}$$

So we have,

$$\frac{\partial}{\partial y_i} \int_{-\infty}^{y} ... \int_{-\infty}^{y_i} ... \int_{-\infty}^{y} \phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C}) dX_n ... dX_i .. dX_1 \Big|_{y_i = y}$$
$$= \frac{\partial}{\partial y_i} \int_{-\infty}^{y} ... \int_{-\infty}^{y_i} ... \int_{-\infty}^{y} \phi(x_1, x_2, ..., x_n; \mathbf{0}, \mathbf{C} \mid X_i = x_i) \phi(x_i; \mathbf{0}, \mathbf{1}) dX_n ... dX_i .. dX_1 \Big|_{y_i = y}$$
$$= \frac{\partial}{\partial y_i} \int_{-\infty}^{y} ... \int_{-\infty}^{y_i} ... \int_{-\infty}^{y} \phi(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n; \boldsymbol{\mu}_{c,i}, \mathbf{C}_{c,i}) \phi(x_i; \mathbf{0}, \mathbf{1}) dX_n ... dX_i .. dX_1 \Big|_{y_i = y}$$
$$= \frac{\partial}{\partial y_i} \int_{-\infty}^{y_i} \phi(x_i; \mathbf{0}, \mathbf{1}) \left(\int_{-\infty}^{y} ... \int_{-\infty}^{y} \phi(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n; \boldsymbol{\mu}_{c,i}, \mathbf{C}_{c,i}) dX_n ... dX_1 \right) dX_i \Big|_{y_i = y}$$
$$= \phi(y; \mathbf{0}, \mathbf{1}) \Phi(y; \boldsymbol{\mu}_{c,i}, \mathbf{C}_{c,i})$$

where we denote $\Phi\left(y;\boldsymbol{\mu}_{c,i},\mathbf{C}_{c,i}\right)=\left(\int_{-\infty}^{y}...\int_{-\infty}^{y}\phi\left(x_1,...,x_{i-1},x_{i+1},...,x_n;\boldsymbol{\mu}_{c,i},\mathbf{C}_{c,i}\right)dX_n...dX_1\right)$.

With a simple transformation to make the conditional distribution with zero means, we get,

$$\Phi\left(y;\boldsymbol{\mu}_{c,i},\mathbf{C}_{c,i}\right)=\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)$$

Thus, the pdf of Y is,

$$f\left(Y=y\right)=\sum_{i=1}^{n}\phi\left(y;\mathbf{0},\mathbf{1}\right)\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)$$

The probability of $Y\geq y$ is calculated as,

$$\begin{aligned}\Pr\left(Y\geq y\right)&=\int_{y}^{\infty}f\left(Y=t\right)dt\\&=\sum_{i=1}^{n}\int_{y}^{\infty}\phi\left(t;\mathbf{0},\mathbf{1}\right)\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)dt\end{aligned}$$

So, we have,

$$\begin{aligned}R&=\Pr\left(Y\geq y\right)\Big/\left(n\left(1-\Phi\left(y\right)\right)\right)\\&=\frac{\sum_{i=1}^{n}\int_{y}^{\infty}\phi\left(t;\mathbf{0},\mathbf{1}\right)\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)dt}{n\left(1-\Phi\left(y\right)\right)}\end{aligned}$$

Taking the derivative of R with respect to y, we obtain,

$$\begin{aligned}\frac{d}{dy}R=&-\frac{n\left(1-\Phi\left(y\right)\right)\sum_{i=1}^{n}\phi\left(y;\mathbf{0},\mathbf{1}\right)\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n^2\left(1-\Phi\left(y\right)\right)^2}\\&-\frac{-n\phi\left(y;\mathbf{0},\mathbf{1}\right)\sum_{i=1}^{n}\int_{y}^{\infty}\phi\left(t;\mathbf{0},\mathbf{1}\right)\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)dt}{n^2\left(1-\Phi\left(y\right)\right)^2}\end{aligned}$$

Because $1-\mathbf{C}_{-i,n}\geq 0$,

$$\begin{aligned}\int_{y}^{\infty}\phi\left(t;\mathbf{0},\mathbf{1}\right)\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)dt&\geq\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\int_{y}^{\infty}\phi\left(t;\mathbf{0},\mathbf{1}\right)dt\\&=\left(1-\Phi\left(y\right)\right)\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\end{aligned}$$

Then, we obtain that,

$$\frac{d}{dy}R \geq -\frac{n\left(1-\Phi(y)\right)\sum_{i=1}^{n}\phi(y;\mathbf{0},\mathbf{1})\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n^2\left(1-\Phi(y)\right)^2} + \frac{n\phi(y;\mathbf{0},\mathbf{1})\sum_{i=1}^{n}\left(1-\Phi(y)\right)\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n^2\left(1-\Phi(y)\right)^2}$$

$$\geq 0$$

Therefore, ***R* is an increasing function w.r.t. *y*.**

Part (2):

Since $\lim_{y\to\infty}\sum_{i=1}^{n}\int_{y}^{\infty}\phi(t;\mathbf{0},\mathbf{1})\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)dt = 0$ and $\lim_{y\to\infty} n\left(1-\Phi(y)\right) = 0$, we apply the L'Hôpital's rule and obtain that,

$$\lim_{y\to\infty} R = \lim_{y\to\infty}\frac{\frac{d}{dy}\sum_{i=1}^{n}\int_{y}^{\infty}\phi(t;\mathbf{0},\mathbf{1})\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)dt}{\frac{d}{dy}\left(n\left(1-\Phi(y)\right)\right)}$$

$$= \lim_{y\to\infty}\frac{-\sum_{i=1}^{n}\phi(y;\mathbf{0},\mathbf{1})\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{-n\phi(y;\mathbf{0},\mathbf{1})}$$

$$= \lim_{y\to\infty}\frac{\sum_{i=1}^{n}\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n}$$

If $\rho_{ij} < 1$, we have $1-\mathbf{C}_{-i,n} > 0$. Then, $\lim_{y\to\infty}\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right) = 1$. Finally, we get,

$$\lim_{y\to\infty} R = \frac{\sum_{i=1}^{n}\lim_{y\to\infty}\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n} = 1$$

Part (3):

Because the correlation coefficient ρ_{ij} is always less than or equal to 1, we have $1-\mathbf{C}_{-i,n} \geq 0$. Thus, $\Phi\left(t\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right) \geq \Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)$ for $t \geq y$. Then we get,

$$R \geq \frac{\sum_{i=1}^{n} \Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\int_{y}^{\infty}\phi(t;\mathbf{0},\mathbf{1})dt}{n\left(1-\Phi(y)\right)} = \frac{\sum_{i=1}^{n} \Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n}$$

We can rewrite that,

$$\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right) = 1-\left(1-\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\right)$$

Note that $\left(1-\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\right)$ is the probability,

$$\begin{aligned}&\left(1-\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\right)\\ &= \Pr\left(\left\{X_1 \geq (1-\rho_{i1})\,y\right\}...\left\{X_{i-1} \geq (1-\rho_{i,i-1})\,y\right\} \cup \left\{X_{i+1} \geq (1-\rho_{i,i+1})\,y\right\}...\left\{X_n \geq (1-\rho_{i,n})\,y\right\} \mid X_i = y\right)\end{aligned}$$

where we assume X_j, $j=1,...,n$ and $j \neq i$ have mean zero and covariance matrix $\mathbf{C}_{c,i}$.

From Boole's inequality we know that,

$$\begin{aligned}&\Pr\left(\left\{X_1 \leq (1-\rho_{i1})\,y\right\}...,\left\{X_{i-1} \leq (1-\rho_{i,i-1})\,y\right\} \cup \left\{X_{i+1} \leq (1-\rho_{i,i+1})\,y\right\}...\left\{X_n \leq (1-\rho_{i,n})\,y\right\} \mid X_i = y\right)\\ &\leq \sum_{j=1,\,j\neq i}^{n} \Pr\left(X_j \leq (1-\rho_{ij})\,y \mid X_i = y\right)\end{aligned}$$

Since the correlation coefficient between X_i and X_j is ρ_{ij}, based on the theory of the bivariate normal distribution, we have that the conditional variance for X_j, given $X_i = y$, is,

$$Var\left(X_j \mid X_i = y\right) = 1-\rho_{ij}^2$$

Actually, the conditional variance can also be obtained by checking the diagonal elements in $\mathbf{C}_{c,i}$, recalling that $\mathbf{C}_{c,i} = \mathbf{C}_{-i,-i} - \mathbf{C}_{-i,n}\mathbf{C}_{n,-i}$.

Then, by standardizing the conditional distribution of X_j given $X_i = y$, we can get that,

$$\Pr\left(X_j \geq (1-\rho_{ij})\,y \mid X_i = y\right) = 1-\Phi\left(y\frac{1-\rho_{ij}}{\sqrt{1-\rho_{ij}^2}}\right) = \Phi\left(-y\sqrt{\frac{1-\rho_{ij}}{1+\rho_{ij}}}\right)$$

Hence,

$$\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right) \geq 1-\sum_{j=1,\,j\neq i}^{n} \Phi\left(-y\times\sqrt{\left(1-\rho_{ij}\right)/\left(1+\rho_{ij}\right)}\right)$$

On the other hand, $\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)$ is probability and should always be non-negative. We have,

$$\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)\geq\max\left(0,1-\sum\nolimits_{j=1,j\neq i}^{n}\Phi\left(-y\times\sqrt{\left(1-\rho_{ij}\right)/\left(1+\rho_{ij}\right)}\right)\right)$$

Finally, we reach that,

$$R\geq\frac{\sum_{i=1}^{n}\Phi\left(y\left(1-\mathbf{C}_{-i,n}\right);\mathbf{0},\mathbf{C}_{c,i}\right)}{n}$$

$$\geq\frac{\sum_{i=1}^{n}\max\left(0,1-\sum_{j=1,j\neq i}^{n}\Phi\left(-y\times\sqrt{\left(1-\rho_{ij}\right)/\left(1+\rho_{ij}\right)}\right)\right)}{n}$$

Q.E.D ■

Appendix 6. *Theorem 4:* $L(\theta)$ is a concave function.

Proof:

Taking the first derivative of $L(\theta)$ with respect to θ, we obtain,

$$\frac{\partial}{\partial\theta}L(\theta)=\frac{\partial}{\partial\theta}\left(\sum_{i=1}^{N}\left(\left(1-l_i\right)\theta^{T}Y_i+l_i\log\left(1-e^{\theta^{T}Y_i}\right)\right)\right)$$

$$=\sum_{i=1}^{N}\left(\left(1-l_i\right)Y_i+l_i\frac{\partial}{\partial\theta}\log\left(1-e^{\theta^{T}Y_i}\right)\right)$$

$$=\sum_{i=1}^{N}\left(\left(1-l_i\right)Y_i+l_i\frac{e^{\theta^{T}Y_i}}{1-e^{\theta^{T}Y_i}}Y_i\right)$$

Taking the second derivative of $L(\theta)$ with respect to θ, we obtain,

$$\frac{\partial^2}{\partial\theta^2}L(\theta)=\frac{\partial}{\partial\theta}\left(\sum_{i=1}^{N}\left(\left(1-l_i\right)Y_i+l_i\frac{e^{\theta^{T}Y_i}}{1-e^{\theta^{T}Y_i}}Y_i\right)\right)$$

$$=\sum_{i=1}^{N}\left(l_i\frac{\partial}{\partial\theta}\left(\frac{e^{\theta^{T}Y_i}}{1-e^{\theta^{T}Y_i}}Y_i\right)\right)$$

$$=-\sum_{i=1}^{N}\left(l_i\frac{e^{\theta^{T}Y_i}}{\left(1-e^{\theta^{T}Y_i}\right)^{2}}Y_iY_i^{T}\right)$$

Because $l_i \geq 0$, $e^{\theta^T Y_i} \Big/ \left(1-e^{\theta^T Y_i}\right)^2 \geq 0$ and $Y_i Y_i^T$ is a positive definite matrix, recalling the property that a positive definite matrix timing a non-negative real number results in a positive definite matrix, $l_i\, e^{\theta^T Y_i} \Big/ \left(1-e^{\theta^T Y_i}\right)^2 Y_i Y_i^T$ is a positive definite matrix. Utilizing another property that the summation of positive definite matrices is a positive definite matrix, we get the conclusion that $-\frac{\partial^2}{\partial \theta^2} L(\theta)$ is a positive definite matrix. Then, $L(\theta)$ is a concave function.

Q.E.D ■

Appendix 7. ***Lemma 5.1:*** **Denote** $\Pr(dis \mid \mathrm{X}) = \Pr(\mathrm{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_M = x_M)$ **and** $\Pr(dis \mid \mathrm{X}_{-i}) = \Pr(\mathrm{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_{i-1} = x_{i-1}, \mathrm{X}_{i+1} = x_{i+1}, ..., \mathrm{X}_M = x_M)$ **. Assume that** $\Pr(\mathrm{X}_i = x_i \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_{i-1} = x_{i-1}, \mathrm{X}_{i+1} = x_{i+1}, ..., \mathrm{X}_M = x_M) = \Pr(\mathrm{X}_i = x_i)$ **. If** $\log\left(1-\Pr(dis \mid \mathrm{X})\right) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$ **, then** $\log\left(1-\Pr(dis \mid \mathrm{X}_{-i})\right) = \beta_0^{'} + \sum_{m=1, m\neq i}^{M} \beta_m x_m$ **, where** $\beta_0^{'} = \beta_0 + \log\left(\sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$

Proof:
We have,

$$\begin{aligned}\Pr(dis \mid \mathrm{X}_{-i}) &= \sum_{\mathrm{X}_i = x_i} \Pr(dis, \mathrm{X}_i \mid \mathrm{X}_{-i}) \\ &= \sum_{\mathrm{X}_i = x_i} \Pr(dis \mid \mathrm{X}_i = x_i, \mathrm{X}_{-i}) \Pr(\mathrm{X}_i = x_i, \mid \mathrm{X}_{-i}) \\ &= \sum_{\mathrm{X}_i = x_i} \Pr(dis \mid \mathrm{X}) \Pr(\mathrm{X}_i = x_i, \mid \mathrm{X}_{-i})\end{aligned}$$

Form $\log\left(1-\Pr(dis \mid \mathrm{X})\right) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$, we get that,

$$\Pr(dis \mid \mathrm{X}) = 1 - e^{\beta_0 + \sum_{m=1}^{M} \beta_m x_m}$$

And from $\Pr(\mathrm{X}_i = x_i \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_{i-1} = x_{i-1}, \mathrm{X}_{i+1} = x_{i+1}, ..., \mathrm{X}_M = x_M) = \Pr(\mathrm{X}_i = x_i)$, we obtain that,

$$\Pr(dis \mid \mathrm{X}_{-i}) = \sum_{x_i} \left(1 - e^{\beta_0 + \sum_{m=1}^{M} \beta_m x_m}\right) \Pr(\mathrm{X}_i = x_i)$$

$$= \sum_{\mathrm{X}_i = x_i} \Pr(\mathrm{X}_i = x_i) - \sum_{\mathrm{X}_i = x_i} e^{\beta_0 + \sum_{m=1}^{M} \beta_m x_m} \Pr(\mathrm{X}_i = x_i)$$

$$= 1 - \sum_{X_i = x_i} e^{\beta_0 + \sum_{m=1, m \neq i}^{M} \beta_m x_m} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)$$

$$= 1 - e^{\beta_0 + \sum_{m=1, m \neq i}^{M} \beta_m x_m} \sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)$$

So we finally get,

$$\log\left(1 - \Pr(dis \mid \mathrm{X}_{-i})\right) = \log\left(e^{\beta_0 + \sum_{m=1, m \neq i}^{M} \beta_m x_m} \sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$$

$$= \beta_0 + \sum_{m=1, m \neq i}^{M} \beta_m x_m + \log\left(\sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$$

$$= \beta_0' + \sum_{m=1, m \neq i}^{M} \beta_m x_m$$

where $\beta_0' = \beta_0 + \log\left(\sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$.

Q.E.D ■

Appendix 8. ***Theorem 5:*** **Denote** $\Pr(dis \mid \mathrm{X}) = \Pr(\mathrm{Status} = disease \mid \mathrm{X}_1 = x_1, ..., \mathrm{X}_M = x_M)$**. Assume that** $\mathrm{X}_1, ..., \mathrm{X}_M$ **are independent to each other. If** $\log\left(1 - \Pr(dis \mid \mathrm{X})\right) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$ **, then for any subset of** $\mathrm{S} \subset \{\mathrm{X}_1, ..., \mathrm{X}_M\}$ **, we have** $\log\left(1 - \Pr(dis \mid \mathrm{S})\right) = \beta_0' + \sum_{X_m \in S} \beta_m x_m$ **, where** $\beta_0' = \beta_0 + \sum_{X_i \in \bar{S}} \log\left(\sum_{X_i = x_i} e^{\beta_i x_i} \Pr(\mathrm{X}_i = x_i)\right)$**.**

Proof:

When the size of S is equal to the size of $\{\mathrm{X}_1, ..., \mathrm{X}_M\}$, that is, $\mathrm{S} = \{\mathrm{X}_1, ..., \mathrm{X}_M\}$, the statement in the theorem trivially holds.

So we assume the size of S is strictly less than that of $\{\mathrm{X}_1, ..., \mathrm{X}_M\}$. Suppose $\bar{\mathrm{S}} = \{X_{k_1}, ..., X_{k_C}\}$, where C is the number of variables not included in S. Then S can be obtained by deleting X_k sequentially. By

writing $S(0)=\{X_1,...,X_M\}$, let us denote $S(j)=S(j-1)\cap\overline{\{X_{k_j}\}}$, for $j=1,...,C$. We can see that $S=S(C)$.

Here, we prove the statement by induction.

When $j=0$, the statement holds simply because $S(0)=\{X_1,...,X_M\}$.

Now assume $\log(1-\Pr(dis\,|\,S(j)))=\beta_0^{'}+\sum_{X_m\in S(j)}\beta_m x_m$ for j, where $\beta_0^{'}=\beta_0+\sum_{X_i\in\{X_{k_1},...,X_{kj}\}}\log\left(\sum_{X_i=x_i}e^{\beta_i x_i}\Pr(X_i=x_i)\right)$.

Then, for $j+1$, from $S(j)=S(j-1)\cap\overline{\{X_{k_j}\}}$ and ***Lemma 5.1*** we have,

$$\log(1-\Pr(dis\,|\,S(j+1)))=\beta_0^{''}+\sum_{X_m\in S(j+1)}\beta_m x_m$$

where $\beta_0^{''}=\beta_0^{'}+\sum_{X_i\in\{X_{k_{j+1}}\}}\log\left(\sum_{X_i=x_i}e^{\beta_i x_i}\Pr(X_i=x_i)\right)=\beta_0+\sum_{X_i\in\{X_{k_1},...,X_{k_{j+1}}\}}\log\left(\sum_{X_i=x_i}e^{\beta_i x_i}\Pr(X_i=x_i)\right)$.

Q.E.D ■

Appendix 9. ***Lemma 6.1:*** **Denote** $\Pr(dis\,|\,X)=\Pr(\text{Status}=disease\,|\,X_1=x_1,...,X_M=x_M)$. **Assume** X_i **is not measured and a correlated variable** $X_i^{'}$ **is observed. Assume the conditional independence that** $\Pr(X_i\,|\,X_1,...,X_{i-1},X_i^{'},X_{i+1},...,X_M)=\Pr(X_i\,|\,X_i^{'})$. **Write** $\Pr(X_i=x_i|X_i^{'}=x_i^{'})=p(x_i^{'},x_i)$. **Denote** Ω_m **the range of** X_m **and** ν_m **the smallest element in** Ω_m, $m=1,...,M$. **Denote** $\Omega_i^{'}$ **the range of** $X_i^{'}$ **and** $\nu_i^{'}$ **the smallest element in** $\Omega_i^{'}$. **If** $\log(1-\Pr(dis\,|\,X))=\beta_0+\sum_{m=1}^{M}\left[\sum_{\omega_m\in\Omega_m\setminus\nu_m}\beta_m(\omega_m)I(\omega_m=x_m)\right]$, **then we have** $\log(1-\Pr(dis\,|\,X_1,...,X_{i-1},X_i^{'},X_{i+1},...,X_M))=\beta_0^{'}+\sum_{m\neq i}\left[\sum_{\omega_m\in\Omega_m\setminus\nu_m}\beta_m(\omega_m)I(\omega_m=x_m)\right]+\sum_{\omega_i^{'}\in\Omega_i^{'}\setminus\nu_i^{'}}\beta_i^{'}(\omega_i^{'})I(\omega_i^{'}=x_i^{'})$.

Proof:

$$\Pr(dis\,|\,X_1,...,X_{i-1},X_i^{'},X_{i+1},...,X_M)$$
$$=\sum_{x_i}\Pr(dis,X_i=x_i\,|\,X_1,...,X_{i-1},X_i^{'},X_{i+1},...,X_M)$$

$$= \sum_{x_i} \Pr\left(dis \mid \mathrm{X}_1,...,\mathrm{X}_{i-1},\mathrm{X}_i = x_i,\mathrm{X}_{i+1},...,\mathrm{X}_M\right) p\left(x_i^{'},x_i\right)$$

$$= \sum_{x_i} \left(1 - e^{\beta_0 + \sum_{m=1}^{M}\left[\sum_{\omega_m \in \Omega_m \setminus v_m} \beta_m(\omega_m) I(\omega_m = x_m)\right]}\right) p\left(x_i^{'},x_i\right)$$

$$= 1 - \sum_{x_i} e^{\beta_0 + \sum_{m=1}^{M}\left[\sum_{\omega_m \in \Omega_m \setminus v_m} \beta_m(\omega_m) I(\omega_m = x_m)\right]} p\left(x_i^{'},x_i\right)$$

$$= 1 - e^{\beta_0 + \sum_{m \neq i}\left[\sum_{\omega_m \in \Omega_m \setminus v_m} \beta_m(\omega_m) I(\omega_m = x_m)\right]} \sum_{x_i} e^{\sum_{\omega_i \in \Omega_i \setminus v_i} \beta_i(\omega_i) I(\omega_i = x_i)} p\left(x_i^{'},x_i\right)$$

Write $b_0 = \sum_{x_i} e^{\sum_{\omega_i \in \Omega_i \setminus v_i} \beta_i(\omega_i) I(\omega_i = x_i)} p\left(v_i^{'},x_i\right)$. For $\omega_i^{'} \in \Omega_i^{'} \setminus v_i^{'}$, write that,

$$\beta_i^{'}\left(\omega_i^{'}\right) = \log\left(\sum_{x_i} e^{\sum_{\omega_i \in \Omega_i \setminus v_i} \beta_i(\omega_i) I(\omega_i = x_i)} p\left(\omega_i^{'},x_i\right)\right) - \log\left(b_0\right).$$

Then, we get that,

$$\log\left(1 - \Pr\left(dis \mid \mathrm{X}_1,...,\mathrm{X}_{i-1},\mathrm{X}_i^{'},\mathrm{X}_{i+1},...,\mathrm{X}_M\right)\right)$$

$$= \beta_0^{'} + \sum_{m \neq i}\left[\sum_{\omega_m \in \Omega_m \setminus v_m} \beta_m\left(\omega_m\right) I\left(\omega_m = x_m\right)\right] + \sum_{\omega_i^{'} \in \Omega_i^{'} \setminus v_i^{'}} \beta_i^{'}\left(\omega_i^{'}\right) I\left(\omega_i^{'} = x_i^{'}\right)$$

where $\beta_0^{'} = \beta_0 + \log\left(b_0\right)$. The above equation can be easily verified by checking the values when $\mathrm{X}_i^{'} = x_i^{'} \in \Omega_i^{'}$.

Q.E.D ■

Appendix 10. ***Theorem 6:*** **X_m, $m = 1,...,M$ are M disease-causing markers but not directly measured. $\mathrm{X}_m^{'}$, $m = 1,...,M$ are M observed markers correlated to X_m, respectively. Assume the conditional independence that $\Pr\left(\mathrm{X}_i \mid \mathrm{X}_1,...,\mathrm{X}_{i-1},\mathrm{X}_i^{'},\mathrm{X}_{i+1},...,\mathrm{X}_M\right) = \Pr\left(\mathrm{X}_i \mid \mathrm{X}_i^{'}\right)$. Denote Ω_m the range of X_m and v_m the smallest element in Ω_m, $m = 1,...,M$. Denote $\Omega_m^{'}$ the range of $\mathrm{X}_m^{'}$ and $v_m^{'}$ the smallest element in $\Omega_m^{'}$, $m = 1,...,M$. If $\log\left(1 - \Pr\left(dis \mid \mathrm{X}\right)\right) = \beta_0 + \sum_{m=1}^{M}\left[\sum_{\omega_m \in \Omega_m \setminus v_m} \beta_m\left(\omega_m\right) I\left(\omega_m = x_m\right)\right]$, then $\log\left(1 - \Pr\left(dis \mid \mathrm{X}\right)\right)$ follows a similar equation and $\log\left(1 - \Pr\left(dis \mid \mathrm{X}^{'}\right)\right) = \beta_0^{'} + \sum_{m=1}^{M}\left[\sum_{\omega_m^{'} \in \Omega_m^{'} \setminus v_m^{'}} \beta_m^{'}\left(\omega_m^{'}\right) I\left(\omega_m^{'} = x_m^{'}\right)\right]$.**

Proof:

If $\mathrm{X}_m^{'} = \mathrm{X}_m$ for $m = 1,...,M$, $\log\left(1-\Pr\left(dis \mid \mathrm{X}^{'}\right)\right) = \beta_0^{'} + \sum_{m=1}^{M}\left[\sum_{\omega_m^{'} \in \Omega_m^{'} \setminus \nu_m^{'}} \beta_m^{'}\left(\omega_m^{'}\right) I\left(\omega_m^{'} = x_m^{'}\right)\right]$ is trivially holds which can be verified by recognizing all parameters are kept unchanged.

So we assume the statement holds with the first K variables $\left(X_k, k \le K\right)$ are replaced by $X_k^{'}, k \le K$,

$$\log\left(1-\Pr\left(dis \mid \mathrm{X}^{'}\right)\right) = \beta_0^{'} + \sum_{m=1}^{M}\left[\sum_{\omega_m^{'} \in \Omega_m^{'} \setminus \nu_m^{'}} \beta_m^{'}\left(\omega_m^{'}\right) I\left(\omega_m^{'} = x_m^{'}\right)\right].$$

Suppose X_{K+1} are not observable and replaced by $\mathrm{X}_{K+1}^{'}$. Then, from the above assumption and from **Lemma 6.1**, we have,

$$\log\left(1-\Pr\left(dis \mid \mathrm{X}^{'}\right)\right) = \beta_0^{''} + \sum_{m=1}^{M}\left[\sum_{\omega_m^{''} \in \Omega_m^{''} \setminus \nu_m^{'}} \beta_m^{''}\left(\omega_m^{''}\right) I\left(\omega_m^{''} = x_m^{''}\right)\right]$$

where all parameters are kept unchanged except for those associated the first variable X_{k+1}. So, the statement holds for $K+1$.

Through induction, the statement is proved.

Q.E.D ■

Appendix 11. ***Corollary 6.1:*** X_m**,** $m = 1,...,M$ **are** M **disease-causing markers but not directly measured.** $\mathrm{X}_m^{'}$**,** $m = 1,...,M$ **are M observed markers correlated to** X_m**, respectively. Assume the conditional independence that** $\Pr\left(\mathrm{X}_i \mid \mathrm{X}_1,...,\mathrm{X}_{i-1},\mathrm{X}_i^{'},\mathrm{X}_{i+1},...,\mathrm{X}_M\right) = \Pr\left(\mathrm{X}_i \mid \mathrm{X}_i^{'}\right)$**. Denote** Ω_m **the range of** X_m **and** ν_m **the smallest element in** Ω_m**,** $m = 1,...,M$ **. If** $\mathrm{X}_m^{'}$ **are all binary with only two possible values** $\{0,1\}$ **and** $\log\left(1-\Pr\left(dis \mid \mathrm{X}\right)\right) = \beta_0 + \sum_{m=1}^{M}\left[\sum_{\omega_m \in \Omega_m \setminus \nu_m} \beta_m\left(\omega_m\right) I\left(\omega_m = x_m\right)\right]$ **, we have,** $\log\left(1-\Pr\left(dis \mid \mathrm{X}^{'}\right)\right) = \beta_0^{'} + \sum_{m=1}^{M} \beta_m^{'} x_m^{'}$ **.**

Proof:

From ***Theorem 6***, we know that,

$$\log\left(1-\Pr\left(dis\mid \mathrm{X}^{'}\right)\right)=\beta_0^{'}+\sum_{m=1}^{M}\left[\sum_{\omega_m^{'}\in\Omega_m^{'}\setminus\nu_m^{'}}\beta_m^{'}\left(\omega_m^{'}\right)I\left(\omega_m^{'}=x_m^{'}\right)\right].$$

Because the size of $\Omega_m^{'}=\{0,1\}$ is 2 and $\nu_m^{'}=0$, $\omega_m^{'}\in\Omega_m^{'}\setminus\nu_m^{'}$ can take value of 1 only. Then, we get,

$$\sum_{\omega_m^{'}\in\Omega_m^{'}\setminus\nu_m^{'}}\beta_m^{'}\left(\omega_m^{'}\right)I\left(\omega_m^{'}=x_m^{'}\right)=\beta_m^{'}(1)I\left(x_m^{'}=1\right)$$

And we can easily verify that,

$$\beta_m^{'}(1)I\left(x_m^{'}=1\right)=\beta_m^{''}x_m^{'}$$

where $\beta_m^{''}=\beta_m^{'}(1)$.

So, we have,

$$\log\left(1-\Pr\left(dis\mid \mathrm{X}^{'}\right)\right)=\beta_0^{'}+\sum_{m=1}^{M}\beta_m^{'}x_m^{'}.$$

Q.E.D ■

Appendix 12. ***Corollary 6.2:*** X_m, $m=1,...,M$ **are** M **disease-causing markers but not directly measured.** $\mathrm{X}_m^{'}$, $m=1,...,M$ **are M observed markers correlated to** X_m, **respectively. If** $\log\left(1-\Pr\left(dis\mid \mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\left[\sum_{\omega_m\in\Omega_m\setminus\nu_m}\beta_m\left(\omega_m\right)I\left(\omega_m=x_m\right)\right]$, **after dominant/recessive genetic coding of** $\mathrm{X}_m^{'}$, **we have,** $\log\left(1-\Pr\left(dis\mid \mathrm{X}^{'}\right)\right)=\beta_0^{'}+\sum_{m=1}^{M}\beta_m^{'}x_m^{'}$.

Proof:

Dominant coding and recessive coding are two special cases of binary $\mathrm{X}_m^{'}$. So, it naturally follows ***Corollary 6.2***.

Q.E.D ■

Appendix 13. ***Theorem 7:*** **If the given disease '** dis **' contains** N **subtypes, that is,** $dis=\bigcup_{n=1}^{N}dis_n$. **Denote X the disease factors under investigation and** $\Pr\left(dis_n\mid \mathrm{X}\right)=\Pr\left(\mathrm{Status}=dis_n\mid \mathrm{X}_1=x_1,...,\mathrm{X}_M=x_M\right)$. **Assume the conditional independence among** dis_n, **that is,**

$\Pr(dis_1,...,dis_N \mid \mathrm{X}) = \prod_{n=1}^{N} \Pr(dis_n \mid \mathrm{X})$. **If** $\log(1-\Pr(dis_n \mid \mathrm{X})) = \beta_{n,0} + \sum_{m=1}^{M} \beta_{n,m} x_m$ **for** $n = 1,...,N$, **then we have** $\log(1-\Pr(dis \mid \mathrm{X})) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$, **where** β_0 **and** β_m **are constants independent to X, and more specifically,** $\beta_0 = \sum_{n=1}^{N} \beta_{n,0}$ **and** $\beta_m = \sum_{n=1}^{N} \beta_{n,m}$.

Proof:

Proof:

Since $dis = \bigcup_{n=1}^{N} dis_n$, we have $\overline{dis} = \bigcap_{n=1}^{N} \overline{dis_n}$. Then, $\Pr(\overline{dis} \mid \mathrm{X}) = \Pr\left(\bigcap_{n=1}^{N} \overline{dis_n} \mid \mathrm{X}\right)$. Because we assume the conditional independence among dis_n, from the definition of independence, the events of being $\overline{dis_n}$ are also independent to each other conditioned on the status of X. Hence, we get,

$$\begin{aligned}\Pr(\overline{dis} \mid \mathrm{X}) &= \prod_{n=1}^{N} \Pr(\overline{dis_n} \mid \mathrm{X}) \\ &= \prod_{n=1}^{N} (1-\Pr(dis_n \mid \mathrm{X}))\end{aligned}$$

Taking the logarithm of both sides, we obtain,

$$\begin{aligned}\log(\Pr(\overline{dis} \mid \mathrm{X})) &= \sum_{n=1}^{N} \log(1-\Pr(dis_n \mid \mathrm{X})) \\ &= \sum_{n=1}^{N} \left(\beta_{n,0} + \sum_{m=1}^{M} \beta_{n,m} x_m\right) \\ &= \sum_{n=1}^{N} \beta_{n,0} + \sum_{n=1}^{N} \left(\sum_{m=1}^{M} \beta_{n,m} x_m\right) \\ &= \sum_{n=1}^{N} \beta_{n,0} + \sum_{m=1}^{M} \left(\sum_{n=1}^{N} \beta_{n,m}\right) x_m \\ &= \beta_0 + \sum_{m=1}^{M} \beta_m x_m\end{aligned}$$

where $\beta_0 = \sum_{n=1}^{N} \beta_{n,0}$ and $\beta_m = \sum_{n=1}^{N} \beta_{n,m}$.

Q.E.D ■

Appendix 14. *Theorem 8:* X_1 and X_2 are two binary variables. Suppose we know $\Pr(dis \mid \mathrm{X}_1 = 0, \mathrm{X}_2 = 0) = p_{00}$, $\Pr(dis \mid \mathrm{X}_1 = 0, \mathrm{X}_2 = 1) = p_{01}$ and $\Pr(dis \mid \mathrm{X}_1 = 1, \mathrm{X}_2 = 0) = p_{10}$, $p_{01} \geq p_{00}$ and $p_{10} \geq p_{00}$. Denote p_{11} the predicted value of $\Pr(dis \mid \mathrm{X}_1 = 1, \mathrm{X}_2 = 1)$ according to model that $\log(\Pr(dis \mid \mathrm{X}) / (1-\Pr(dis \mid \mathrm{X}))) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$. Denote $p_{11}^{'}$ the predicted value of $\Pr(dis \mid \mathrm{X}_1 = 1, \mathrm{X}_2 = 1)$ according to model that $\log(1-\Pr(dis \mid \mathrm{X})) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$. We have $p_{11} \geq p_{11}^{'}$, where the equality holds if and only if $p_{01} = p_{00}$ or $p_{10} = p_{00}$.

Proof:

From $\log\left(\Pr\left(dis\,|\,\mathrm{X}\right)/\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$, we have,

$$\frac{p_{11}}{1-p_{11}}=\frac{\frac{p_{01}}{1-p_{01}}\times\frac{p_{10}}{1-p_{10}}}{\frac{p_{00}}{1-p_{00}}}=\frac{p_{01}p_{10}\left(1-p_{00}\right)}{\left(1-p_{01}\right)\left(1-p_{10}\right)p_{00}}.$$

With simple algebra operations, we get,

$$1-p_{11}=\frac{\left(1-p_{01}\right)\left(1-p_{10}\right)p_{00}}{p_{01}p_{10}\left(1-p_{00}\right)}p_{11}$$

and

$$p_{11}=\frac{p_{01}p_{10}\left(1-p_{00}\right)}{p_{00}-p_{01}p_{00}-p_{10}p_{00}+p_{01}p_{10}}.$$

From $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$, we have,

$$p_{11}^{'}=1-\frac{\left(1-p_{01}\right)\left(1-p_{10}\right)}{1-p_{00}}.$$

since,

$$\begin{aligned}1-p_{11}^{'}&=\left(1-p_{00}\right)\times\left(\frac{1-p_{10}}{1-p_{00}}\right)\times\left(\frac{1-p_{01}}{1-p_{00}}\right)\\&=\frac{\left(1-p_{10}\right)\left(1-p_{01}\right)}{1-p_{00}}\end{aligned}$$

where we have utilized the facts that $\beta_0=\log\left(1-p_{00}\right)$, $\beta_1=\log\left(\frac{1-p_{10}}{1-p_{00}}\right)$ and $\beta_2=\log\left(\frac{1-p_{01}}{1-p_{00}}\right)$.

Now, let us check the sign of $p_{11}-p_{11}^{'}$. We have,

$$\begin{aligned}p_{11}-p_{11}^{'}&=\left(1-p_{11}^{'}\right)-\left(1-p_{11}\right)\\&=\left(1-p_{11}\right)\left(\frac{1-p_{11}^{'}}{1-p_{11}}-1\right)\\&=\left(1-p_{11}\right)\left(\frac{p_{01}p_{10}}{p_{00}p_{11}}-1\right)\end{aligned}$$

Here, we would like to check the value of $p_{01}p_{10}/(p_{00}p_{11})$ compared to 1.

$$\frac{p_{01}p_{10}}{p_{00}p_{11}} = \frac{p_{00} - p_{00}p_{01} - p_{00}p_{10} + p_{01}p_{10}}{p_{00}(1-p_{00})}$$

$$= \frac{p_{01}(p_{10} - p_{00}) + p_{00} - p_{00}p_{10}}{p_{00}(1-p_{00})}$$

$$= \frac{(p_{01} - p_{00})(p_{10} - p_{00}) + p_{00}(p_{10} - p_{00}) + p_{00} - p_{00}p_{10}}{p_{00}(1-p_{00})}$$

$$= \frac{(p_{01} - p_{00})(p_{10} - p_{00}) + p_{00}(1-p_{00})}{p_{00}(1-p_{00})}$$

Because $p_{01} - p_{00} \geq 0$ and $p_{10} - p_{00} \geq 0$, we have,

$$\frac{p_{01}p_{10}}{p_{00}p_{11}} \geq \frac{p_{00}(1-p_{00})}{p_{00}(1-p_{00})} = 1.$$

And since $1 - p_{11} > 0$, finally we prove that,

$$p_{11} - p_{11}^{'} \geq 0.$$

If $p_{01} = p_{00}$ or $p_{10} = p_{00}$, we get $(p_{01} - p_{00})(p_{10} - p_{00}) = 0$ and $p_{11} = p_{11}^{'}$.

If $p_{11} = p_{11}^{'}$, we can easily get $(p_{01} - p_{00})(p_{10} - p_{00}) = 0$. So $p_{01} = p_{00}$ or $p_{10} = p_{00}$.

Q.E.D ■

Appendix 15. ***Corollary 8.1:*** **$X_1 \in \{0,1\}$ and $X_2 \in \{0,1\}$ are two binary variables with '1' denoting the disease-risk increasing genotypes. Suppose X_1 and X_2 have interactions and $\Pr(dis \mid X_1 = 1, X_2 = 1)$ is larger than the value predicted by logistic regression model. The detection power through logistic regression is smaller than the power through the model $\log(1 - \Pr(dis \mid X)) = \beta_0 + \sum_{m=1}^{M} \beta_m x_m$.**

Proof:

To detect the interaction effects, we add one interaction term to both logistic regression model and the model $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$. Thus, both two models have four parameters. Recalling that for binary variables X_1 and X_2 there are totally four genotypes, both two models have the same log-likelihoods, which is denoted by L_H .

Denote L_L the log-likelihood associated with the logistic regression model without interaction term and denote $L_L^{'}$ the log-likelihood associated with the model $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$.

Denote p_{11} the predicted value of $\Pr\left(dis\,|\,\mathrm{X}_1=1,\mathrm{X}_2=1\right)$ according to the logistic regression model that $\log\left(\Pr\left(dis\,|\,\mathrm{X}\right)/\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$. Denote $p_{11}^{'}$ the predicted value of $\Pr\left(dis\,|\,\mathrm{X}_1=1,\mathrm{X}_2=1\right)$ according to model that $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$. For the genotype $\left(\mathrm{X}_1=1,\mathrm{X}_2=1\right)$, from ***Theorem 8*** we know that $p_{11}\geq p_{11}^{'}$. At the same time, we know $\Pr\left(dis\,|\,\mathrm{X}_1=1,\mathrm{X}_2=1\right)\geq p_{11}$. So the logistic regression model will have a better prediction of $\Pr\left(dis\,|\,\mathrm{X}_1=1,\mathrm{X}_2=1\right)$.

Thus,

$$L_L\geq L_L^{'}.$$

The log likelihood ratio test calculates the chi-square statistics for the logistic regression as

$$\chi^2=2\left(L_H-L_L\right).$$

For the model $\log\left(1-\Pr\left(dis\,|\,\mathrm{X}\right)\right)=\beta_0+\sum_{m=1}^{M}\beta_m x_m$, similarly we have the log-likelihood ratio test as,

$$\chi^{2'}=2\left(L_H-L_L^{'}\right).$$

So we can easily get that $\chi^{2'}\geq\chi^2$, which indicates that the hypothesis testing through logistic regression model will have a larger p-value.

Q.E.D ■

Appendix 16. ***Theorem 9:*** **The expected number of TGEP is an invariant quantity. Assume g_1 and g_2 are two mutually disjoint genotypes, of which g_1 a disease-risk genotype is and g_2 is not a disease-risk genotype. Denote $\Upsilon(g_1)$ and $\Upsilon(g_1 \cup g_2)$ the estimated numbers based on g_1 and the union of g_1 and g_2, respectively. Then, $E(\Upsilon(g_1)) = E(\Upsilon(g_1 \cup g_2))$.**

Proof:

Denote j and M the numbers of cases and total subjects, respectively. Let k_1 and k_2 represent the numbers of cases having the genotypes g_1 and g_2, respectively. Similarly, r_1 and r_2 stands for the numbers of subjects specified by the genotypes g_1 and g_2.

Based on the genotype g_1, the number of TGEP is estimated by,

$$\Upsilon(g_1) = k_1 - \frac{j - k_1}{M - r_1} r_1$$

Because g_1 and g_2 are disjoint, the cases and subjects associated with the genotypes union $g_1 \cup g_2$ can be calculated directly as, $k_1 + k_2$ and $r_1 + r_2$, respectively. Based on the union $g_1 \cup g_2$, the number of TGEP is estimated by,

$$\Upsilon(g_1 \cup g_2) = (k_1 + k_2) - \frac{j - k_1 - k_2}{M - r_1 - r_2}(r_1 + r_2)$$

Since g_2 is not a disease-risk genotype and g_1 is a disease-risk genotype, given the condition of g_1, k_2 follows the hyper-geometric distribution with parameter triplet $(M - r_1, j - k_1, r_2)$. Then, we have the expectation $E(k_2 \mid r_1, k_1)$ as,

$$E(k_2 \mid r_1, k_1) = \frac{j - k_1}{M - r_1} r_2$$

Thus, we have the expectation $E(\Upsilon(g_1 \cup g_2))$ as follows,

$$E\left(\Upsilon\left(g_1 \cup g_2\right)\right) = E\left(\left(k_1 + k_2\right) - \frac{j - k_1 - k_2}{M - r_1 - r_2}\left(r_1 + r_2\right)\right)$$

$$= E\left(k_1 + E\left(k_2 \mid r_1, k_1\right) - \frac{j - k_1 - E\left(k_2 \mid r_1, k_1\right)}{M - r_1 - r_2}\left(r_1 + r_2\right)\right)$$

$$= E\left(k_1 + \frac{j - k_1}{M - r_1} r_2 - \frac{j - k_1 - \frac{j - k_1}{M - r_1} r_2}{M - r_1 - r_2}\left(r_1 + r_2\right)\right)$$

$$= E\left(k_1 + \frac{j - k_1}{M - r_1} r_2 - \frac{j - k_1}{M - r_1}\left(r_1 + r_2\right)\right)$$

$$= E\left(k_1 - \frac{j - k_1}{M - r_1} r_1\right)$$

Recall that we have the expectation for $\Upsilon\left(g_1\right)$ as,

$$E\left(\Upsilon\left(g_1\right)\right) = E\left(k_1 - \frac{j - k_1}{M - r_1} r_1\right)$$

Hence, we obtain that,

$$E\left(\Upsilon\left(g_1 \cup g_2\right)\right) = E\left(\Upsilon\left(g_1\right)\right)$$

Q.E.D ■

Appendix 17. *Corollary 9.1:* Assume the genotype *g* is the only disease-risk genotype specified by SNPs subset S containing *n* interacting SNPs, the number of TGEP associated with *g* can be estimated by its lower order genotype specified by a proper subset of S.

Proof:

In this proof, we call a genotype specified by *d* SNPs as a *d*-SNP genotype. Assuming an *m*-SNPs genotype g' is *g*'s lower order genotype, then, g' is specified by $1 \le m \le d-1$ SNPs and the *m* SNPs have exactly the same values as in the genotype *g*. Under the flexible dominant/recessive model, the genotype g' can be represented by the union of 2^{d-m} *d*-SNPs genotypes,

$$g' = g_0 \cup g_1 \cup g_2 \cup ... \cup g_{2^{d-m}-1}$$

These 2^{d-m} *d*-SNPs genotypes have *m* SNPs' values the same as in the *d*-SNP genotype *g*. The combination of the remaining $d-m$ SNPs generates totally 2^{d-m} possible *d*-SNPs genotypes. Since *g* is an *d*-SNPs genotype, *g* must be one of the 2^{d-m} *d*-SNPs genotypes. For the sake of discussion we denote that $g_0 = g$. Because *g* is the only disease-risk genotype specified by SNPs subset S , the *d*-SNP genotype $g_{i\neq 0}$ is not a disease-risk genotype. On the other hand, we know these 2^{d-m} *d*-SNPs genotypes are mutually disjoint. Thus, considering the genotypes g_0 and g_1, according to ***theorem 9***, we can obtain that,

$$E\left(\Upsilon\left(g_0 \cup g_1\right)\right) = E\left(\Upsilon\left(g_0\right)\right).$$

Now, considering $g_0 \cup g_1$ as whole and adding in g_2 , and reapplying ***theorem 9***, we get,

$$E\left(\Upsilon\left(g_0 \cup g_1 \cup g_2\right)\right) = E\left(\Upsilon\left(g_0 \cup g_1\right)\right) = E\left(\Upsilon\left(g_0\right)\right)$$

Iteratively adding in one more *d*-SNPs genotype, we will finally have that,

$$E\left(\Upsilon\left(g^{'}\right)\right) = E\left(\Upsilon\left(g_0 \cup g_1 \cup ... \cup g_{2^{d-m}-1}\right)\right) = E\left(\Upsilon\left(g_0\right)\right)$$

Hence, it is reasonable to estimate the number of TGEP associated with *g* through its lower order genotype $g^{'}$.

Q.E.D ■

Appendix 18. ***Theorem 10:*** **Assume g_1 and g_2 are genotypes specified by two non-overlapping SNP subsets. Suppose there are totally M subjects with j subjects being cases. The genotype g_1 specifies a population with r_1 subjects and k_1 cases. The genotype g_2 specifies a population with r_2 subjects and k_2 cases. Write $\alpha = j/M$, $B_1 = k_1/r_1$ and $B_2 = k_2/r_2$. The maximum likelihood estimation of the number of TGEP associated with $g_1 \cap g_2$ is $\frac{\sigma_1^2\Upsilon_1 + \sigma_2^2\Upsilon_2}{\sigma_1^2 + \sigma_2^2}$ and the variance of the estimation is $\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$, where $\Upsilon_1 = k_1 - \frac{j-k_1}{M-r_1}r_1$, $\Upsilon_2 = k_2 - \frac{j-k_2}{M-r_2}r_2$,**

$\sigma_1^2 = \frac{M(1-B_1B_2)\alpha_1(1-\alpha_1)B_1^{'}(1-B_1^{'})}{(1-B_1)^2}$, $\sigma_2^2 = \frac{M(1-B_1B_2)\alpha_2(1-\alpha_2)B_2^{'}(1-B_2^{'})}{(1-B_2)^2}$, $\alpha_1 = \alpha - \frac{\Upsilon_1}{M}$, $\alpha_2 = \alpha - \frac{\Upsilon_2}{M}$,

$B_1^{'} = \frac{B_1(1-B_2)}{1-B_1B_2}$, **and** $B_2^{'} = \frac{B_2(1-B_1)}{1-B_1B_2}$.

Proof:

From the definition of TGEP and equation (24), we have the estimated numbers of TGEP associated with the genotypes g_1 and g_2 as,

$$\Upsilon_1 = k_1 - \frac{j-k_1}{M-r_1}r_1$$

$$\Upsilon_2 = k_2 - \frac{j-k_2}{M-r_2}r_2 .$$

Denote the number of TGEP associated with $g_1 \cap g_2$ as Υ. Since $g_1 \cap g_2$ is considered the disease-risk genotype and $g_1 = (g_1 \cap g_2) \cup (g_1 \cap \overline{g_2})$, according to ***Theorem 9***, we have,

$$\Upsilon_1 = \Upsilon + \varepsilon_1$$

where ε_1 can be considered as a zero-mean Gaussian noise. Here, we show the variance of ε_1 is $\sigma_1^2 = \frac{M(1-B_1B_2)\alpha_1(1-\alpha_1)B_1^{'}(1-B_1^{'})}{(1-B_1)^2}$ where $\alpha_1 = \alpha - \frac{\Upsilon_1}{M}$ and $B_1^{'} = \frac{B_1(1-B_2)}{1-B_1B_2}$.

Suppose there are r subjects and k cases carrying the genotype $g_1 \cap g_2$. Similarly, we assume there are $r^{'}$ subjects and $k^{'}$ cases carrying the genotype $g_1 \cap \overline{g_2}$. Because $g_1 = (g_1 \cap g_2) \cup (g_1 \cap \overline{g_2})$, we have $r_1 = r + r^{'}$ and $k_1 = k + k^{'}$. Then, we can rewrite Υ_1 as,

$$\begin{aligned}\Upsilon_1 &= k + k^{'} - \frac{j-k-k^{'}}{M-r_1}r_1 \\ &= k - \frac{j-k}{M-r_1}r_1 + k^{'} + \frac{k^{'}}{M-r_1}r_1 \\ &= k - \frac{j-k}{M-r_1}r_1 + \left(\frac{1}{1-B_1}\right)k^{'}\end{aligned}$$

Since we are interested in the genotype $g_1 \bigcap g_2$, k can be considered as constant and the variance associated with Υ_1 comes from the random variable $k^{'}$. $k^{'}$ can be looked as a hypergeometric random variable with r subjects and k cases fixed, that is, $k^{'}$ is randomly drawn from a cell of size $r^{'}$ with total $M-r$ subjects and $j-k$ cases. However, we do not know the values of k, r and $r^{'}$. They can be approximated as $k \approx \Upsilon_1 + \alpha_1 r$, $r = MB_1B_2$ and $r^{'} = MB_1(1-B_2)$. With simple operations we have the approximation for the variance of $k^{'}$ as,

$$\mathrm{var}(k^{'}) \approx M(1-B_1B_2)\alpha_1(1-\alpha_1)B_1^{'}(1-B_1^{'})$$

where $\alpha_1 = \frac{j-k}{M-r} \approx \alpha - \frac{\Upsilon_1}{M}$ and $B_1^{'} = \frac{r^{'}}{M-r} \approx \frac{B_1(1-B_2)}{1-B_1B_2}$.

Then, the variance of ε_1 (equivalent to the variance of Υ_1) is,

$$\sigma_1^2 = \frac{\mathrm{var}(k^{'})}{(1-B_1)^2} = \frac{M(1-B_1B_2)\alpha_1(1-\alpha_1)B_1^{'}(1-B_1^{'})}{(1-B_1)^2}$$

Similarly, we have,

$$\Upsilon_2 = \Upsilon + \varepsilon_2$$

where ε_2 can be considered as a zero-mean Gaussian noise and the variance of ε_2 is $\sigma_2^2 = \frac{M(1-B_1B_2)\alpha_2(1-\alpha_2)B_2^{'}(1-B_2^{'})}{(1-B_2)^2}$ where $\alpha_2 = \alpha - \frac{\Upsilon_2}{M}$ and $B_2^{'} = \frac{B_2(1-B_1)}{1-B_1B_2}$.

Since $(g_1 \bigcap \overline{g_2}) \bigcap (\overline{g_1} \bigcap g_2) = \phi$, ε_1 and ε_2 are independent to each other. Thus, Υ_1 and Υ_2 can be viewed as two independent noisy observations of Υ, our object is to estimate the underlying Υ from the two observations Υ_1 and Υ_2. We derive the estimation based on the maximum likelihood principle.

The log likelihood of observing Υ_1 and Υ_2 with Υ considered as a parameter is,

$$L(\Upsilon_1, \Upsilon_2; \Upsilon) = -\log(2\pi\sigma_1\sigma_2) - \frac{(\Upsilon_1 - \Upsilon)^2}{2\sigma_1^2} - \frac{(\Upsilon_2 - \Upsilon)^2}{2\sigma_2^2}$$

Taking the derivative with respect to Υ, we get,

$$\frac{\partial}{\partial \Upsilon} L = \frac{\Upsilon_1 - \Upsilon}{\sigma_1^2} + \frac{\Upsilon_2 - \Upsilon}{\sigma_2^2}$$

$$= \frac{\Upsilon_1}{\sigma_1^2} + \frac{\Upsilon_2}{\sigma_2^2} - \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \Upsilon$$

Let the derivative equal to zero, we obtain the maximum likelihood estimation as,

$$\hat{\Upsilon} = \frac{\frac{\Upsilon_1}{\sigma_1^2} + \frac{\Upsilon_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\sigma_2^2 \Upsilon_1 + \sigma_1^2 \Upsilon_2}{\sigma_2^2 + \sigma_1^2}$$

The variance of the estimation is,

$$\mathrm{var}\left(\hat{\Upsilon} \right) = \frac{\sigma_2^4 \mathrm{var}\left(\Upsilon_1 \right) + \sigma_1^4 \mathrm{var}\left(\Upsilon_2 \right)}{\left(\sigma_2^2 + \sigma_1^2 \right)^2} = \frac{\sigma_2^4 \sigma_1^2 + \sigma_1^4 \sigma_2^2}{\left(\sigma_2^2 + \sigma_1^2 \right)^2} = \frac{\sigma_2^2 \sigma_1^2}{\sigma_2^2 + \sigma_1^2}$$

Q.E.D

■

Appendix 19. *Theorem 11:* Assume $\Phi(\bullet)$ is the CDF of the standard Gaussian distribution. Let $f(x) = \prod_{i=1}^{d-1} \left(1 - \left(\Phi\left(\left(\sqrt{2} \right)^{i-1} x \right) \right)^{d-i+1} \right)$. Then, $f(x)$ is a strictly decreasing function and $f^{-1}(\bullet)$ can be solved in logarithm complexity by binary search.

Proof:

Denote $\varphi(\bullet)$ the PDF of the standard Gaussian distribution. Taking the derivative of $f(x)$, we get,

$$\frac{\partial}{\partial x} f(x) = -\sum_{m=1}^{d-1} \left(\left(\Phi\left(\left(\sqrt{2} \right)^{m-1} x \right) \right)^{d-m} \varphi\left(\left(\sqrt{2} \right)^{i-1} x \right) \left(\sqrt{2} \right)^{i-1} \prod_{i \neq m} \left(1 - \left(\Phi\left(\left(\sqrt{2} \right)^{i-1} x \right) \right)^{d-i+1} \right) \right)$$

Because $\left(\Phi\left(\left(\sqrt{2} \right)^{m-1} x \right) \right)^{d-m} > 0$, $\varphi\left(\left(\sqrt{2} \right)^{i-1} x \right) > 0$ and $\prod_{i \neq m} \left(1 - \left(\Phi\left(\left(\sqrt{2} \right)^{i-1} x \right) \right)^{d-i+1} \right) > 0$, each item in the summation is strictly larger than 0. Noticing there is negative sign on the left of the summation, we conclude that $\frac{\partial}{\partial x} f(x) < 0$, so $f(x)$ is strictly decreasing.

Given a value h such that $0 < h < 1$, let $x_{01} = \Phi^{-1}\left(\sqrt[d]{1-\sqrt[d-1]{h}}\right)$ and $x_{02} = \Phi^{-1}\left(\sqrt{1-\sqrt[d-1]{h}}\right)$. Then, calculate x_1 and x_2 as the following,

$$x_1 = \begin{cases} x_{01} & \textbf{if } x_{01} > 0 \\ x_{01} \Big/ \left(\sqrt{2}\right)^{d-1} & \textbf{if } x_{01} \le 0 \end{cases}$$

$$x_2 = \begin{cases} x_{02} \Big/ \left(\sqrt{2}\right)^{d-1} & \textbf{if } x_{02} < 0 \\ x_{02} & \textbf{if } x_{02} \le 0 \end{cases}$$

Now we prove that x_1 and x_2 can serve as the two initial points for the binary search to iteratively solve $f^{-1}(h)$, by proving $f(x_1) \le h \le f(x_2)$.

$$\begin{aligned} f(x_1) &= \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1} x_1\right)\right)^{d-i+1}\right) \\ &\le \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1} x_1\right)\right)^{d}\right) \end{aligned}$$

If $x_{01} > 0$, we have,

$$\begin{aligned} & \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1} x_1\right)\right)^{d}\right) \\ &= \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1} x_{01}\right)\right)^{d}\right) \\ &< \prod_{i=1}^{d-1}\left(1-\left(\Phi\left(x_{01}\right)\right)^{d}\right) \\ &= \prod_{i=1}^{d-1}\left(1-\left(\sqrt[d-1]{h}\right)^{d}\right) \\ &= h \end{aligned}$$

If $x_{01} \le 0$, we have,

$$\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1}x_1\right)\right)^d\right)$$

$$=\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-d}x_{01}\right)\right)^d\right)$$

$$<\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(x_{01}\right)\right)^d\right)$$

$$=\prod_{i=1}^{d-1}\left(1-\left(\sqrt[d]{1-\sqrt[d-1]{h}}\right)^d\right)$$

$$=h$$

Thus, $f\left(x_1\right)\le h$.

On the other hand, we have,

$$f\left(x_2\right)=\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1}x_2\right)\right)^{d-i+1}\right)$$

$$\ge\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(\left(\sqrt{2}\right)^{i-1}x_1\right)\right)^2\right)$$

$$\ge\prod_{i=1}^{d-1}\left(1-\left(\Phi\left(x_{02}\right)\right)^2\right)$$

$$=h$$

Thus, $f\left(x_2\right)\ge h$.

Recalling that $f\left(x\right)$ is a strictly decreasing function, we can do the binary search with the two initial points as x_1 and x_2. And we know that the binary search has a logarithm complexity.

Q.E.D ■

Appendix 20. ***Theorem 12:*** **Assume P_d and P_{d-1} be the thresholds of experimental-wise significance for the *d*-order and (*d*-1) order interactions, respectively. Suppose S^d is a significant *d*-**

order interacting SNP subset and g_d is the disease-risk genotype specified by S^d. Denote $F=\left(\Phi^{-1}\left(P_d\right)/\Phi^{-1}\left(P_{d-1}\right)\right)^2$, where $\Phi^{-1}(\bullet)$ is the inverse CDF of standard normal distribution. If the single SNP *s* is included in S^d with *B* being the frequency of disease-risk allele and the (*d*-1)-order subsets of S^d excluding *s* is not significant and, the frequency of the disease risk genotype g_d is larger than $B_{\min}=\left(BF-1\right)/\left(F-1\right)$.

Proof:

Denote Υ as the number of TGEP associated with the genotype g_d. Denote M as the number of total subjects of which α fraction are cases. Let $B^{'}$ denote the frequency of g_d. From the equation (27), we have the *z* score for the *d*-order SNP subset as,

$$\frac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{1}{B^{'}}-1}.$$

Due to the invariant property of Υ, we have the *z* score for the (*d*-1)-order SNP subset of S^d excluding ***s*** as,

$$\frac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\frac{B}{B^{'}}-1}.$$

Here, we assume the frequency $B^{'}$ of g_d is equal to the multiplication of the frequency B and the frequency of the (*d*-1)-order genotype, since $g_d=g\bigcap g_{d-1}$, where g is the risk genotype specified by *s* and g_{d-1} is the (*d*-1)-order genotype specified by the (*d*-1)-order SNP subset of S^d excluding ***s.***

Because the *d*-order SNP subset is significant but the (d-1)-order one is not, we have,

$$\begin{cases}\dfrac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\dfrac{1}{B^{'}}-1}\geq-\Phi^{-1}\left(P_d\right)\\ \dfrac{\Upsilon}{\sqrt{M\alpha(1-\alpha)}}\sqrt{\dfrac{B}{B^{'}}-1}<-\Phi^{-1}\left(P_{d-1}\right)\end{cases}$$

With simple algebraic operations, we can obtain that,

$$\frac{\sqrt{\frac{1}{B^{'}}-1}}{\sqrt{\frac{B}{B^{'}}-1}} \geq \frac{\Phi^{-1}\left(P_d\right)}{\Phi^{-1}\left(P_{d-1}\right)}$$

$$\Rightarrow \sqrt{\frac{1-B^{'}}{B-B^{'}}} \geq \frac{\Phi^{-1}\left(P_d\right)}{\Phi^{-1}\left(P_{d-1}\right)}$$

$$\Rightarrow \frac{1-B^{'}}{B-B^{'}} \geq \left(\frac{\Phi^{-1}\left(P_d\right)}{\Phi^{-1}\left(P_{d-1}\right)}\right)^2 = F$$

Because $g_d = g \bigcap g_{d-1}$, $B \geq B^{'}$. Then, we have,

$$1-B^{'} \geq F\left(B-B^{'}\right)$$

$$\Rightarrow B^{'} \geq \frac{FB-1}{F-1}$$

Q.E.D ■

Appendix 21. ***Theorem 13:*** **Denote X_{ij} the *j*th random variable in the *i*th replication with the p-values p_{ij} associated with each random variable. Assume X_{ij} has identically distribution. Writing $\Pr\left(s=1 \mid p \leq \alpha/M\right)$, we have $\mathbf{E}\left(F_B\right) = \alpha \times \Pr\left(s=1 \mid p \leq \alpha/M\right)$.**

Proof:

From the definition of F_B,

$$F_B = \frac{\sum_{i=1}^{N} T_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} s_{ij} \mathrm{I}\left(p_{ij} \leq \frac{\alpha}{M}\right)}{N},$$

we have that,

$$\mathbf{E}\left(F_B\right) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} \mathbf{E}\left(s_{ij} \mathrm{I}\left(p_{ij} \leq \frac{\alpha}{M}\right)\right)}{N}.$$

Because,

$$\begin{aligned}\mathbf{E}\left(s_{ij}\mathrm{I}\left(p_{ij}\le\frac{\alpha}{M}\right)\right)&=\mathbf{E}\left(s_{ij}\left|p_{ij}\le\frac{\alpha}{M}\right.\right)\Pr\left(p_{ij}\le\frac{\alpha}{M}\right)\\&=\mathbf{E}\left(s_{ij}\left|p_{ij}\le\frac{\alpha}{M}\right.\right)\times\frac{\alpha}{M}\\&=\left(\Pr\left(s_{ij}=1\left|p_{ij}\le\frac{\alpha}{M}\right.\right)\times1+\Pr\left(s_{ij}=1\left|p_{ij}\le\frac{\alpha}{M}\right.\right)\times0\right)\times\frac{\alpha}{M}\\&=\Pr\left(s_{ij}=1\left|p_{ij}\le\frac{\alpha}{M}\right.\right)\times\frac{\alpha}{M}\end{aligned}$$

Since we assume X_{ij} is identically distributed, we get,

$$\mathbf{E}\left(s_{ij}\mathrm{I}\left(p_{ij}\le\frac{\alpha}{M}\right)\right)=\Pr\left(s=1\left|p\le\frac{\alpha}{M}\right.\right)\times\frac{\alpha}{M}$$

Finally, we obtain that,

$$\begin{aligned}\mathbf{E}\left(F_B\right)&=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{E}\left(s_{ij}\mathrm{I}\left(p_{ij}\le\frac{\alpha}{M}\right)\right)}{N}\\&=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\Pr\left(s=1\left|p\le\frac{\alpha}{M}\right.\right)\times\frac{\alpha}{M}}{N}\\&=\alpha\times\Pr\left(s=1\,|\,p\le\alpha/M\right)\end{aligned}$$

Q.E.D. ■

Appendix 22. *Corollary 13.1:* The correlation in the multiple tests does not contribute to the conservativeness in the sense that $\mathbf{E}\left(F_B\right)=\alpha$ if the tests are exhaustively searched.

Proof:

Because the tests are exhaustively searched, we have

$$\Pr\left(s=1\,|\,p\le\alpha/M\right)=1.$$

Hence, from ***Theorem 13*** we get,

$$\mathbf{E}(F_B) = \alpha \times \Pr(s=1 \mid p \le \alpha/M) = \alpha .$$

Q.E.D. ■

Appendix 23. *Theorem 14:* For large number of independent multiple tests, $\mathbf{E}(F_A) \approx \frac{1-\mathbf{e}^{-c\alpha}}{c\alpha}\mathbf{E}(F_B)$, where α is the Bonferroni corrected significance level and $c = \Pr(s=1 \mid p \le \alpha/M)$. Especially, when $\alpha = 0.05$ and $c = 1$, $\mathbf{E}(F_A) \approx 0.98\,\mathbf{E}(F_B)$.

Proof:

Denote M the number of multiple tests and T_i the number of false positives in the *i*th replication dataset. Recalling that,

$$F_A = \frac{\sum_{i=1}^{N} \mathrm{I}(T_i \ge 1)}{N}.$$

Then, we have,

$$\mathbf{E}(F_A) = \frac{\sum_{i=1}^{N} \left(1 - \Pr(T_i = 0)\right)}{N}.$$

Since $T_i = \sum_{j=1}^{M} s_{ij} \mathrm{I}\left(p_{ij} \le \frac{\alpha}{M}\right)$, we can write that,

$$\Pr(T_i = 0) = \Pr\left(\left(\overline{s_{i1} = 1 \cap p_{i1} \le \alpha/M}\right) \cap \ldots \cap \left(\overline{s_{iM} = 1 \cap p_{iM} \le \alpha/M}\right)\right)$$

$$= \left(1 - \frac{c\alpha}{M}\right)^{M}$$

where $c = \Pr(s=1 \mid p \le \alpha/M)$.

From the theory of limitation, we have,

$$\lim_{M \to \infty} \left(1 - \frac{c\alpha}{M}\right)^{M} = e^{c\alpha}.$$

If M is large enough, we could the following approximation,

$$\left(1-\frac{c\alpha}{M}\right)^{M} \approx e^{c\alpha}.$$

Further, we have from ***Theorem 13*** that,

$$\mathbf{E}(F_B) = c\alpha.$$

So,

$$\mathbf{E}(F_A) \approx \frac{1-\mathbf{e}^{-c\alpha}}{c\alpha}\mathbf{E}(F_B).$$

When $\alpha = 0.05$ and $c = 1$, we can easily calculate that,

$$\frac{1-\mathbf{e}^{-c\alpha}}{c\alpha} = \frac{1-e^{-0.05}}{0.05} = \frac{0.0488}{0.05} \approx 0.98$$

Q.E.D. ■

Appendix 24. *Theorem 15:* If P_G and P_S are two positively dependent variables, $\Pr(P_G \le v | P_S \ge \tau) \le \Pr(P_G \le v)$.

Proof:

From the property of the positively dependent variables, we have that,

$$\Pr(P_G \le v, P_S \le \tau) \ge \Pr(P_G \le v) \times \Pr(P_S \le \tau).$$

Then, we can derive that,

$$\begin{aligned}\Pr(P_G \le v, P_S \ge \tau) &= \Pr(P_G \le v) - \Pr(P_G \le v, P_S \le \tau) \\ &\le \Pr(P_G \le v) - \Pr(P_G \le v) \times \Pr(P_S \le \tau) \\ &= \Pr(P_G \le v)\left(1 - \Pr(P_S \le \tau)\right) \\ &= \Pr(P_G \le v) \times \Pr(P_S \ge \tau)\end{aligned}$$

So,

$$\Pr\left(P_G \le v \mid P_S \ge \tau\right) = \frac{\Pr\left(P_G \le v, P_S \ge \tau\right)}{\Pr\left(P_S \ge \tau\right)} \le \Pr\left(P_G \le v\right)$$

Q.E.D. ■

Bibliography

1. Agresti, A. (2002) *Categorical data analysis*. Wiley-Interscience, New York.
2. Allen, N.E., *et al.* (2009) Moderate alcohol intake and cancer incidence in women, *J Natl Cancer Inst*, **101**, 296-305.
3. Bakay, M., *et al.* (2006) Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration, *Brain*, **129**, 996-1013.
4. Boyd, S.P. and Vandenberghe, L. (2004) *Convex optimization*. Cambridge University Press, Cambridge, UK ; New York.
5. Braga-Neto, U.M. and Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification?, *Bioinformatics (Oxford, England)*, **20**, 374-380.
6. Cai, Z., *et al.* (2007) Selecting dissimilar genes for multi-class classification, an application in cancer subtyping, *BMC Bioinformatics*, **8**, 206.
7. Castellsague, X., *et al.* (1999) Independent and joint effects of tobacco smoking and alcohol drinking on the risk of esophageal cancer in men and women, *Int J Cancer*, **82**, 657-664.
8. Chen, L., *et al.* (2011) Comparative analysis of methods for detecting interacting loci, *BMC Genomics*, **12**, 344.
9. Clarke, R., *et al.* (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, *Nat Rev Cancer*, **8**, 37-49.
10. Cormen, T.H. and Cormen, T.H. (2001) *Introduction to algorithms*. MIT Press, Cambridge, Mass.
11. Cover, T.M. and Thomas, J.A. (2006) *Elements of information theory*. Wiley-Interscience, Hoboken, N.J.
12. Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern classification*. Wiley, New York.

13. Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77-87.
14. Duerr, R.H., *et al.* (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene, *Science*, **314**, 1461-1463.
15. Durbin, B.P., *et al.* (2002) A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics*, **18 Suppl 1**, S105-110.
16. Durrett, R. (1996) *Probability theory and examples*. Duxbury press.
17. Fort, G. and Lambert-Lacroix, S. (2005) Classification using partial least squares with penalized logistic regression, *Bioinformatics*, **21**, 1104-1111.
18. Garcia-Closas, M., *et al.* (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses, *Lancet*, **366**, 649-659.
19. Garro, A.J. and Lieber, C.S. (1990) Alcohol and cancer, *Annu Rev Pharmacol Toxicol*, **30**, 219-249.
20. Genz, A. (1992) Numerical Computation of Multivariate Normal Probabilities, *Journal of Computational and Graphical Statistics*, **1**.
21. Gish, H. (1990) A probabilistic approach to the understanding and training of neural network classifiers. *IEEE Intl. Conf. Acoust., Speech, Signal Process.*, Alburquerque, NM, pp. 1361-1364.
22. Goldstein, D.B. (2009) Common genetic variation and human traits, *N Engl J Med*, **360**, 1696-1698.
23. Golub, T.R., *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.
24. Gu, J., *et al.* (2005) Effects of N-acetyl transferase 1 and 2 polymorphisms on bladder cancer risk in Caucasians, *Mutat Res*, **581**, 97-104.
25. Guyon, I., *et al.* (2002) Gene selection for cancer classification using support vector machines, *Machine Learning*, **46**, 389-422.
26. Hanczar, B. and Dougherty, E.R. (2010) On the Comparison of Classifiers for Microarray Data, *Current Bioinformatics*, **5**, 29-39.

27. Harley, J.B., *et al.* (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci, *Nat Genet*, **40**, 204-210.
28. Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York.
29. Hindorff, L.A., *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proc Natl Acad Sci U S A*, **106**, 9362-9367.
30. Huber, W., *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, **18 Suppl 1**, S96-104.
31. Knuth, D.E. (1997) *The art of computer programming*. Addison-Wesley, Reading, Mass.
32. Lai, C., *et al.* (2006) A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets, *BMC Bioinformatics*, **7**, 235.
33. Lee, C.H., *et al.* (2005) Independent and combined effects of alcohol intake, tobacco smoking and betel quid chewing on the risk of esophageal cancer in Taiwan, *Int J Cancer*, **113**, 475-482.
34. Legnani, C., *et al.* (2002) Venous thromboembolism in young women; role of thrombophilic mutations and oral contraceptive use, *Eur Heart J*, **23**, 984-990.
35. Lewis, S.J. and Smith, G.D. (2005) Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach, *Cancer Epidemiol Biomarkers Prev*, **14**, 1967-1971.
36. Li, F. and Yang, Y. (2005) Analysis of recursive gene selection approaches from microarray data, *Bioinformatics*, **21**, 3741-3747.
37. Li, T., Zhang, C. and Ogihara, M. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*, **20**, 2429-2437.
38. Liu, H., Li, J. and Wong, L. (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, *Genome Informatics*, **13**, 51-60.
39. Liu, J.J., *et al.* (2005) Multiclass cancer classification and biomarker discovery using GA-based algorithms, *Bioinformatics*, **21**, 2691-2697.

40. Loog, M., Duin, R.P.W. and Haeb-Umbach, R. (2001) Multiclass linear dimension reduction by weighted pairwise Fisher criteria, *IEEE Trans Pattern Anal Machine Intell*, **23**, 762-766.
41. Maher, B. (2008) Personal genomes: The case of the missing heritability, *Nature*, **456**, 18-21.
42. Manolio, T.A., *et al.* (2009) Finding the missing heritability of complex diseases, *Nature*, **461**, 747-753.
43. Marchini, J., Donnelly, P. and Cardon, L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases, *Nat Genet*, **37**, 413-417.
44. Martinelli, I., *et al.* (1999) Interaction between the G20210A mutation of the prothrombin gene and oral contraceptive use in deep vein thrombosis, *Arterioscler Thromb Vasc Biol*, **19**, 700-703.
45. Matsuo, K., *et al.* (2001) Gene-environment interaction between an aldehyde dehydrogenase-2 (ALDH2) polymorphism and alcohol consumption for the risk of esophageal cancer, *Carcinogenesis*, **22**, 913-916.
46. McCarthy, M.I., *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nat Rev Genet*, **9**, 356-369.
47. McCullagh, P. and Nelder, J.A. (1989) *Generalized linear models*. Chapman and Hall, London ; New York.
48. McLachlan, G.J. and Krishnan, T. (2008) *The EM algorithm and extensions*. Wiley-Interscience, Hoboken, N.J.
49. Moore, J.H., *et al.* (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, *J Theor Biol*, **241**, 252-261.
50. Park, M.Y. and Hastie, T. (2008) Penalized logistic regression for detecting gene interactions, *Biostatistics*, **9**, 30-50.
51. Pasternak, J.J. (2005) *An introduction to human molecular genetics : mechanisms of inherited diseases*. Wiley-Liss, Hoboken, N.J.
52. Pomeroy, S.L., *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, **415**, 436-442.
53. Ramaswamy, S., *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures, *Proc Natl Acad Sci U S A*, **98**, 15149-15154.

54. Rifkin, R. and Klautau, A. (2002) In Defense of One-Vs-All classification, *Journal of Machine Learning Research*, **5**, 101-141.
55. Ritchie, M.D., *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *Am J Hum Genet*, **69**, 138-147.
56. Rocke, D.M. and Durbin, B. (2001) A model for measurement error for gene expression arrays, *J Comput Biol*, **8**, 557-569.
57. Rosendaal, F.R. (1999) Venous thrombosis: a multicausal disease, *Lancet*, **353**, 1167-1173.
58. Rosing, J., *et al.* (1999) Low-dose oral contraceptives and acquired resistance to activated protein C: a randomised cross-over study, *Lancet*, **354**, 2036-2040.
59. Ruczinski, I., Kooperberg, C. and Leblanc, M. (2003) Logic regression, *Journal of Computational and Graphical Statistics*, **12**, 475-511.
60. Sachidanandam, R., *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature*, **409**, 928-933.
61. Sanderson, S., Salanti, G. and Higgins, J. (2007) Joint effects of the N-acetyltransferase 1 and 2 (NAT1 and NAT2) genes and smoking on bladder carcinogenesis: a literature-based systematic HuGE review and evidence synthesis, *Am J Epidemiol*, **166**, 741-751.
62. Schena, M., *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science (New York, N.Y.*, **270**, 467-470.
63. Scott, L.J., *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants, *Science*, **316**, 1341-1345.
64. Seligsohn, U. and Lubetsky, A. (2001) Genetic susceptibility to venous thrombosis, *N Engl J Med*, **344**, 1222-1231.
65. Shedden, K.A., *et al.* (2003) Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework, *Am J Pathol*, **163**, 1985-1995.
66. Shi, L., *et al.* (2008) The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies, *BMC Bioinformatics*, **9 Suppl 9**, S10.
67. Stacey, S.N., *et al.* (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer, *Nat Genet*, **39**, 865-869.

68. Statnikov, A., *et al.* (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, **21**, 631-643.
69. Staunton, J.E., *et al.* (2001) Chemosensitivity prediction by transcriptional profiling, *Proc Natl Acad Sci U S A*, **98**, 10787-10792.
70. Syvanen, A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms, *Nat Rev Genet*, **2**, 930-942.
71. Thorisson, G.A., *et al.* (2005) The International HapMap Project Web site, *Genome research*, **15**, 1592-1593.
72. Tibshirani, R., *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proc Natl Acad Sci U S A*, **99**, 6567-6572.
73. Tipping, M.E. and Bishop, C.M. (1999) Probabilistic Principle Component Analysis, *Journal of the Royal Statistical Society. Series B*, **61**, 611-622.
74. Tong, Y.L. (1990) *The multivariate normal distribution*. Springer series in statistics. Springer-Verlag, New York.
75. Vandenbroucke, J.P., *et al.* (1994) Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation, *Lancet*, **344**, 1453-1457.
76. Vapnik, V.N. (1998) *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York.
77. Wang, Y., *et al.* (2002) Iterative normalization of cNDA microarray data, *IEEE Trans Info. Tech. Biomed*, **6**, 29-37.
78. Wang, Y., Miller, D.J. and Clarke, R. (2008) Approaches to working in high-dimensional data spaces: gene expression microarrays, *Br J Cancer*, **98**, 1023-1028.
79. Wang, Y., *et al.* (2003) Partially-independent component analysis for tissue heterogeneity correction in microarray gene expression analysis. *IEEE Workshop on Neural Networks for Signal Processing*. Toulouse, France, pp. 24-32.
80. Wang, Z., *et al.* (2006) Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data, *Bioinformatics*, **22**, 755-761.
81. Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses, *The Annals of Mathematical Statistics*, **9**, 60-62.
82. Xuan, J., *et al.* (2007) Gene selection for multiclass prediction by weighted fisher criterion, *EURASIP J Bioinform Syst Biol*, 64628.

83. Yang, C., *et al.* (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies, *Bioinformatics*, **25**, 504-511.
84. Yeang, C.H., *et al.* (2001) Molecular classification of multiple tumor types, *Bioinformatics*, **17 Suppl 1**, S316-322.
85. Zeegers, M.P., *et al.* (2000) The impact of characteristics of cigarette smoking on urinary tract cancer risk: a meta-analysis of epidemiologic studies, *Cancer*, **89**, 630-639.
86. Zhang, J., *et al.* (2008) Pattern expression non-negative matrix factorization: Algorithm and application to blind source separation, *Computational Intelligence and Neuroscience*, Artical ID 168769.
87. Zhang, Y. and Liu, J.S. (2007) Bayesian inference of epistatic interactions in case-control studies, *Nat Genet*, **39**, 1167-1173.
88. Zhao, Y., Li, M.C. and Simon, R. (2005) An adaptive method for cDNA microarray normalization, *BMC Bioinformatics*, **6**, 28.
89. Zhou, X. and Tuck, D.P. (2007) MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data, *Bioinformatics*, **23**, 1106-1114.