Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

Evaluating Methods for Modeling Epistasis Networks with Application to Head and Neck Cancer

Rajesh Talluri[1] and Sanjay Shete[1,2]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

ABSTRACT: Epistasis helps to explain how multiple single-nucleotide polymorphisms (SNPs) interact to cause disease. A variety of tools have been developed to detect epistasis. In this article, we explore the strengths and weaknesses of an information theory approach for detecting epistasis and compare it to the logistic regression approach through simulations. We consider several scenarios to simulate the involvement of SNPs in an epistasis network with respect to linkage disequilibrium patterns among them and the presence or absence of main and interaction effects. We conclude that the information theory approach more efficiently detects interaction effects when main effects are absent, whereas, in general, the logistic regression approach is appropriate in all scenarios but results in higher false positives. We compute epistasis networks for SNPs in the *FSD1L* gene using a two-phase head and neck cancer genome-wide association study involving 2,185 cases and 4,507 controls to demonstrate the practical application of the methods.

KEYWORDS: epistasis, head and neck cancer, information theory, networks, regression

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Talluri and Shete. Evaluating Methods for Modeling Epistasis Networks with Application to Head and Neck Cancer. *Cancer Informatics* 2015:14(S2) 17–23 doi: 10.4137/CIN.S17289.

RECEIVED: November 16, 2014. **RESUBMITTED:** January 05, 2015. **ACCEPTED FOR PUBLICATION:** January 06, 2015.

ACADEMIC EDITOR: J.T Efird, Editor in Chief

TYPE: Original Research

FUNDING: This work was supported in part by the National Institutes of Health (grants R01CA131324, R01DE022891, and R25 DA026120 to SS), a cancer prevention fellowship for RT supported by a grant from the National Institute of Drug Abuse (NIH R25 DA026120), and the Barnhart Family Distinguished Professorship in Targeted Therapy (to SS). Additional funding information is in the Acknowledgments section at the end of this paper. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CANCER INFORMATICS: This manuscript evaluates methods for modeling epistasis networks in a head and neck cancer genome-wide association study.

CORRESPONDENCE: sshete@mdanderson.org

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Genome-wide association studies (GWAS) are used to identify single-nucleotide polymorphisms (SNPs) associated with complex diseases such as cancer.[1] However, most GWAS analyze the main effects of SNPs. Epistasis is observed when the effect of an SNP is modified by other SNPs.[2–4] Epistasis between SNPs helps to explain how multiple SNPs interact to cause disease. For example, epistasis between genes has been associated with hypertension,[5] sporadic breast cancer,[6] and several other diseases.[7] Epistasis also plays a subtle part in explaining missing heritability.[8,9] Thus, identifying epistatic SNP interactions is of interest to better understand disease etiology. Furthermore, some studies suggest that, if the epistatic variance is larger than the additive variance, more power can be achieved to detect SNPs by searching for epistasis between SNPs rather than evaluating only the main effects.[10]

A variety of tools have been used to detect epistasis, such as regression,[11–14] Bayesian methods,[15–20] and artificial intelligence algorithms.[21–27] For higher order interactions, where regression methods are not suitable, several machine learning methods such as multifactor dimensionality reduction,[28] tree-based methods,[25] and entropy-based methods[23,29] have been proposed, as they use classifiers and feature selection to reduce the computational burden.

In this article, we use simulations to explore the strengths and weaknesses of an information theory approach[29] for

detecting epistasis compared to the logistic regression approach. We perform studies in which we simulate SNPs with and without the main effects. We also consider three types of interaction patterns and two types of linkage disequilibrium patterns. Finally, we demonstrate the practical application of these methods to identify an epistasis network. We use data from a head and neck cancer GWAS of the *FSD1L* gene that involves 1,154 cases and 1,542 controls. We then attempt to replicate our findings in an independent head and neck cancer GWAS of the *FSD1L* gene that involves 1,031 cases and 2,965 controls.

Materials and Methods

We used a case–control study design to introduce the approaches to epistasis network analysis; however, the methods are also applicable to continuous phenotypes. The case–control status is denoted by a binary indicator Y, which takes the value of 1 or 0, corresponding to the categorization of the individual as being among the cases or the controls. The epistasis networks are networks in which the nodes are SNPs and the edges between the nodes correspond to the interaction between the SNPs. Hereafter, we define the two approaches for developing epistasis networks.

Information theory approach. For ease of presentation, we consider epistasis between two SNPs, A and B. Each SNP can have three possible genotypes: AA, Aa, and aa, which are coded as 0, 1, and 2, respectively, and where a is the minor allele. In the information theory approach, the association of the disease with an SNP or with the interaction between a pair of SNPs is quantified by assigning weights referred to as mutual information when a single SNP is studied and information gain when the interaction between SNPs is studied.[30] In the regression framework, these weights correspond to the respective odds ratios of the main or interaction effects. Specifically, mutual information between two variables provides a measure of the reduction in randomness in a variable when information about another variable is available. The mutual information of SNP A and the case–control status Y (the main effect of SNP A) is defined as

$$I(A;Y) = H(Y) - H(Y|A),$$

where $H(Y)$ is the entropy of Y, which is defined as

$$H(Y) = -\sum_{i\in\{0,1\}} P(Y=i)\log(P(Y=i))$$

and $H(Y|A)$ is the conditional entropy, which is defined as

$$H(Y|A) = -\sum_{i\in\{0,1\}}\sum_{G\in\{AA,Aa,aa\}} P(Y=i,A=G)\log(P(Y=i|A=G))$$

where $P(Y=i|A=G) = \dfrac{P(Y=i,A=G)}{P(A=G)}$.

The mutual information $I(A;Y)$ ranges from 0 to 1. A zero value for the mutual information indicates independence, ie, SNP A has no effect on disease status Y. A higher value of the mutual information indicates a stronger relationship between SNP A and the disease status.

Given a pair of SNPs A and B, the information gain of A and B (interaction effect) is defined as

$$IG(A;B;Y) = I(A,B;Y) - I(A;Y) - I(B;Y).$$

The information gain takes values between –1 and 1. A positive value indicates interactions that explain a part of the phenotypic variance; a zero value indicates interactions that do not explain any phenotypic variance; and a negative value indicates that modeling the interactions will be redundant because the information is already contained in the main effects (ie, modeling would possibly lead to multicollinearity). In this analysis, an information gain greater than zero was considered to represent a significant interaction.

Logistic regression approach. In standard logistic regression modeling, interactions between SNP A and SNP B are evaluated by testing the significance of β_{AB}:

$$logit(P(Y=1)) = \beta_0 + \beta_A\, SNPA + \beta_B\, SNPB + \beta_{AB}\, SNPA * SNPB$$

We used Bonferroni correction to account for multiple comparisons. For an epistasis network of k SNPs, the number of multiple comparisons is the sum of the total number of main effects (k) and the total number of interactions ($k(k-1)/2$). The Bonferroni-corrected P-value used was 0.05/(total number of interactions + total number of main effects).

Simulations. We performed simulation studies to investigate the performance of the methods. We considered several scenarios to simulate the SNPs involved in an epistasis network: scenarios with different linkage disequilibrium patterns and scenarios with presence or absence of main and interaction effects. In scenarios 1 and 2, all the SNPs were in linkage equilibrium, whereas in scenarios 3 and 4 the SNPs were in linkage disequilibrium. We used the linkage disequilibrium pattern of the *FSD1L* gene from the head and neck cancer GWAS data to mimic realistic linkage disequilibrium patterns. In scenarios 1 and 3, all the SNPs were simulated with only interaction effects and without main effects, whereas in scenarios 2 and 4 the SNPs were simulated with both interaction and main effects. For each scenario, we used 10 SNPs to simulate three different epistasis networks (see Figs. 1A, 2A, and 3A, and Table 1). We used a logistic regression model to simulate 10,000 cases and 10,000 control samples:

$$\log it(P(Y=1)) = \beta_0 + \sum_{i=1}^{10}\beta_i SNP\, i + \sum_{ij|i\neq j}\beta_{ij}\, SNP\, i\, SNP\, j$$

where $\beta_0 = -2$. For the different simulation scenarios, the SNPs and interacting pairs that were significant are listed in Table 1.

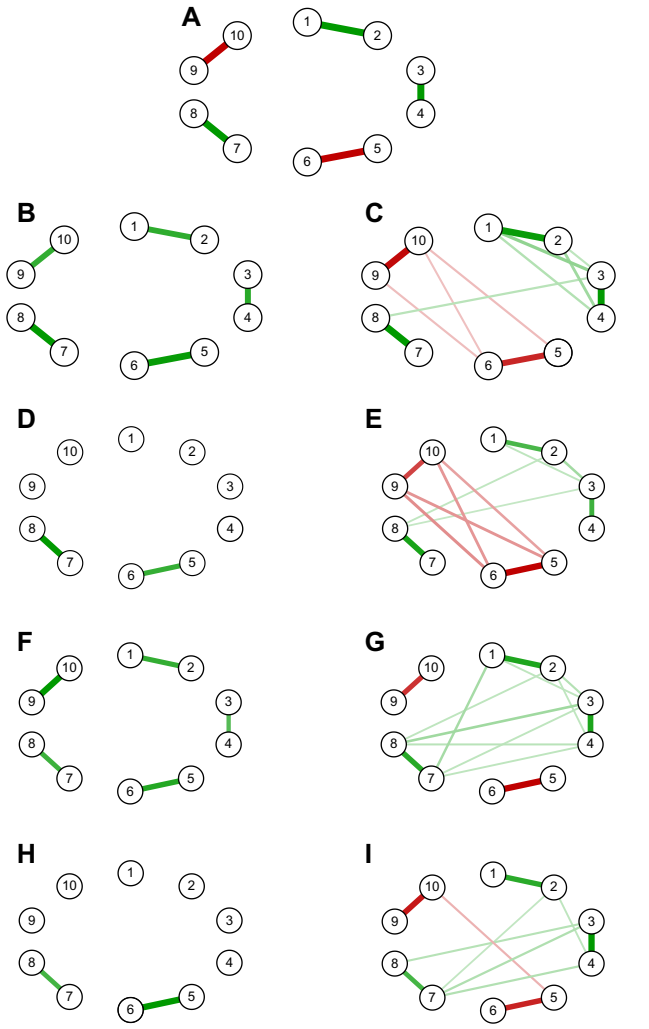

Figure 1. Epistasis networks for the four scenarios simulated on the basis of network 1. (**A**) The true simulated epistasis network. (**B**) Epistasis network for simulation scenario 1 – information theory approach. (**C**) Epistasis network for simulation scenario 1 – logistic regression approach. (**D**) Epistasis network for simulation scenario 2 – information theory approach. (**E**) Epistasis network for simulation scenario 2 – logistic regression approach. (**F**) Epistasis network for simulation scenario 3 – information theory approach. (**G**) Epistasis network for simulation scenario 3 – logistic regression approach. (**H**) Epistasis network for simulation scenario 4 – information theory approach. (**I**) Epistasis network for simulation scenario 4 – logistic regression approach.

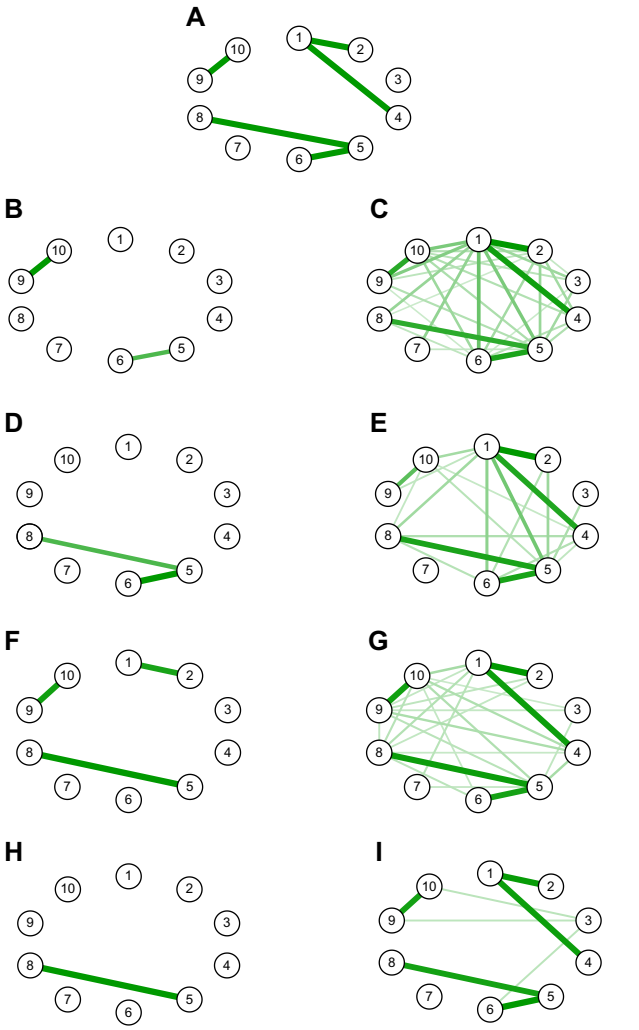

Figure 2. Epistasis networks for the four scenarios simulated on the basis of network 2. (**A**) The true simulated epistasis network. (**B**) Epistasis network for simulation scenario 1 – information theory approach. (**C**) Epistasis network for simulation scenario 1 – logistic regression approach. (**D**) Epistasis network for simulation scenario 2 – information theory approach. (**E**) Epistasis network for simulation scenario 2 – logistic regression approach. (**F**) Epistasis network for simulation scenario 3 – information theory approach. (**G**) Epistasis network for simulation scenario 3 – logistic regression approach. (**H**) Epistasis network for simulation scenario 4 – information theory approach. (**I**) Epistasis network for simulation scenario 4 – logistic regression approach.

Results

We analyzed the simulated data from the four simulation scenarios using the information theory approach and the logistic model approach as described previously. For each simulation scenario, the results are presented for the three interaction networks in Figures 1A, 2A, and 3A (referred to as networks 1, 2, and 3, respectively).

Simulation scenario 1. In simulation scenario 1, the SNPs were in linkage equilibrium and had interacting effects, but no main effects. For the simulation based on network 1 (Fig. 1A), the information theory approach exactly identified all five interaction effects without any false positives (Fig. 1B). The logistic regression approach also identified all five simulated interaction effects; however, it also falsely identified several interactions that were not simulated (Fig. 1C). In the simulation using network 2 (Fig. 2A), which included two SNPs (SNP 5 and SNP 1) that were common in two independent interactions, the information theory approach identified only two of the five interactions simulated (Fig. 2B), whereas the logistic regression approach identified all five interactions; however, it also identified several false positive interactions (Fig. 2C). In the simulation using network 3 (Fig. 3A), which involved only the interaction between SNP 1 and SNP 2,

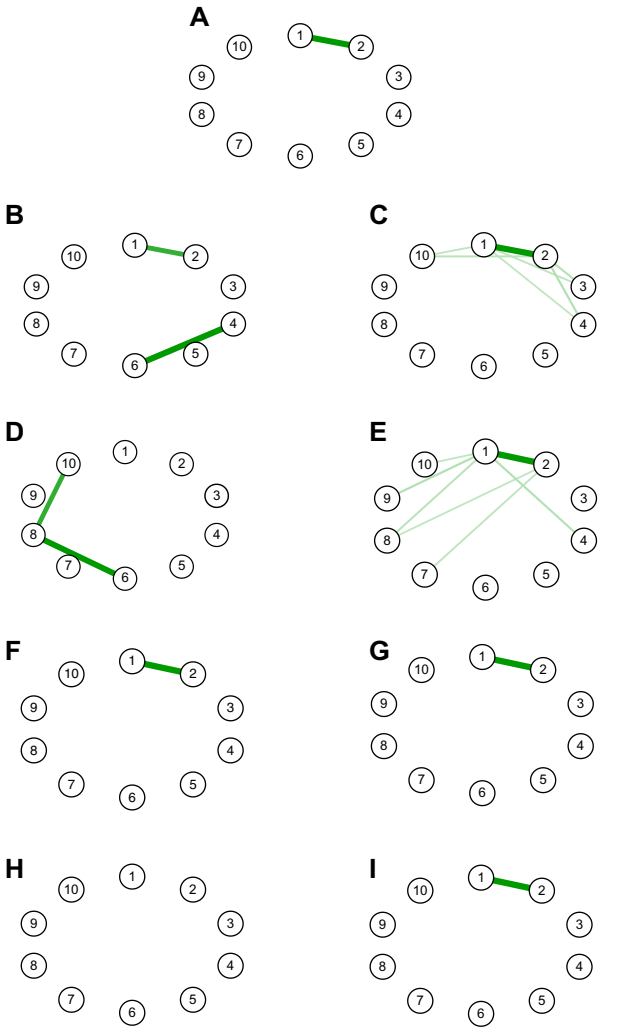

Figure 3. Epistasis networks for the four scenarios simulated on the basis of network 3. (**A**) The true simulated epistasis network. (**B**) Epistasis network for simulation scenario 1 – information theory approach. (**C**) Epistasis network for simulation scenario 1 – logistic regression approach. (**D**) Epistasis network for simulation scenario 2 – information theory approach. (**E**) Epistasis network for simulation scenario 2 – logistic regression approach. (**F**) Epistasis network for simulation scenario 3 – information theory approach. (**G**) Epistasis network for simulation scenario 3 – logistic regression approach. (**H**) Epistasis network for simulation scenario 4 – information theory approach. (**I**) Epistasis network for simulation scenario 4 – logistic regression approach.

Table 1. Details of the four simulation scenarios.

SIMULATION SCENARIO	MAIN EFFECTS	INTERACTION EFFECTS	LINKAGE DISEQUILIBRIUM
Scenario 1	None	(1,2), (3,4), (5,6), (7,8), (9,10)	No
	None	(1,2), (1,4), (5,6), (5,8), (9,10)	No
	None	(1,2)	No
Scenario 2	1, 3, 9	(1,2), (3,4), (5,6), (7,8), (9,10)	No
	1, 3, 9	(1,2), (1,4), (5,6), (5,8), (9,10)	No
	1, 3, 9	(1,2)	No
Scenario 3	None	(1,2), (3,4), (5,6), (7,8), (9,10)	Yes
	None	(1,2), (1,4), (5,6), (5,8), (9,10)	Yes
	None	(1,2)	Yes
Scenario 4	1, 3, 9	(1,2), (3,4), (5,6), (7,8), (9,10)	Yes
	1, 3, 9	(1,2), (1,4), (5,6), (5,8), (9,10)	Yes
	1, 3, 9	(1,2)	Yes

Notes: All the main effects and interaction effects that were present were simulated with an odds ratio of 2.

the information theory approach and the logistic regression approach identified the true simulated interaction; however, both approaches also identified a few false positive interactions (Fig. 3B and C).

Simulation scenario 2. In, simulation scenario 2, all the SNPs were in linkage equilibrium and had both interaction and main effects. For the simulation based on network 1 (Fig. 1A), the information theory approach identified only two of the five interaction effects that were simulated (Fig. 1D). In contrast, the logistic regression approach identified all the simulated interactions; however, it additionally identified several false positives (Fig. 1E). In the simulation using network 2, the information theory approach identified only two of the five interactions simulated (Fig. 2D), whereas the logistic regression approach identified the simulated interactions as well as several false positives (Fig. 2E). In the simulation with network 3, the information theory approach failed to identify the true simulated interaction and identified two false positive interactions (Fig. 3D). In contrast, the logistic regression approach identified the true simulated interaction (Fig. 3E); however, it also identified several false positive interactions.

Simulation scenario 3. In simulation scenario 3, the SNPs were in linkage disequilibrium and had interaction effects, but no main effects. In the simulation using network 1, the information theory approach exactly identified all five interaction effects that were simulated (Fig. 1F), whereas the logistic regression approach identified interactions that were not simulated in addition to the simulated interactions (Fig. 1G). In the simulation using network 2, the information theory approach identified three of the five true simulated interactions, whereas the logistic regression approach identified several false positives in addition to the simulated interactions (Fig. 2F and G). In the simulation using network 3, both approaches identified the true simulated interaction without any false positives (Fig. 3F and G).

Simulation scenario 4. In simulation scenario 4, the SNPs were in linkage disequilibrium and had both inter-

action effects and main effects. For the simulation based on network 1, the information theory approach identified only two of the five interaction effects that were simulated (Fig. 1H), whereas the logistic regression approach identified several false positives in addition to the five simulated interactions (Fig. 1I). For the simulation based on network 2, the information theory approach identified only one of the five true simulated interactions, whereas the logistic regression approach identified several false positives in addition to the five simulated interactions (Fig. 2H and I). For the simulation based on network 3, the information theory approach failed to identify the true simulated interaction, whereas the logistic regression approach identified the true simulated interaction (Fig. 3H and I).

Head and neck cancer data. We applied both approaches to data from a GWAS of head and neck cancer. The study participants were patients at The University of Texas MD Anderson Cancer Center (UT MD Anderson) with newly diagnosed, histologically confirmed, previously untreated head and neck cancer, including cancers of the oral cavity, pharynx, and larynx. The study genotyping was performed in two phases. The data from phase 1 included 2,696 individuals: 1,154 head and neck cancer patients and 1,542 controls. The data from phase 2 included 3,996 individuals: 1,031 cases and 2,965 controls. The institutional review board at UT MD Anderson approved the case–control study, and all participants provided written informed consent.

In this analysis, we developed epistasis networks for SNPs within the *FSD1L* gene. The *FSD1L* gene is located on chromosome 9 and is mainly expressed in neural tissue. The *FSD1L* gene codes for type 2 cystatins, which regulate the activity of endogenous cysteine proteinases such as cathepsin B, H, S, L, and K. These enzymes are involved in tumor cell invasion and metastasis.[31] Therefore, we hypothesized that interacting SNPs in this gene may play a role in head and neck cancer etiology. In our study, a total of 617 SNPs were genotyped in the *FSD1L* gene. However, some of the SNPs were in high linkage disequilibrium. Our simulation study showed that linkage disequilibrium was confounded with epistasis (simulation study data not shown). Therefore, we considered only the SNPs in this gene locus that were in low linkage disequilibrium ($r^2 < 0.1$) to develop the epistasis network for head and neck cancer.

We computed the epistasis network for the phase 1 data and used the phase 2 data to validate the epistasis network. The epistasis networks we developed for the phase 1 data by using the information theory approach and the logistic regression approach are shown in Figure 4A and B, respectively. The epistasis network based on the information theory approach identified the interaction between SNP *rs*630103 and SNP *rs*10122572 to be significant, whereas the epistasis network based on the logistic regression approach identified the interaction between two different SNPs to be significant, namely, SNP *rs*2812312 and SNP *rs*2049347. The epistasis networks we developed for the phase 2 data (the validation dataset) by using the information theory approach and the logistic regression approach are shown in Figure 5A and B, respectively. In the validation dataset, the information theory approach identified that the interactions between SNPs rs2049347, *rs*7038470, and *rs*10122572 are associated with head and neck cancer. The logistic regression approach identified that the interaction between SNP *rs*2812312 and SNP *rs*10990985 is significantly associated with head and neck cancer. None of the interactions identified from the phase 1 epistasis networks was replicated in the phase 2 epistasis networks.

Discussion

In this paper, we compare the information theory approach and the logistic regression approach for modeling epistasis networks. We used simulations to explore the strengths and weaknesses of the two approaches. We considered several simulation scenarios to simulate SNPs involved in an epistasis network with varying degrees of linkage disequilibrium patterns and the presence or absence of main and interaction effects.

The information theory approach accurately identified the epistasis network when there were no main effects. However, in

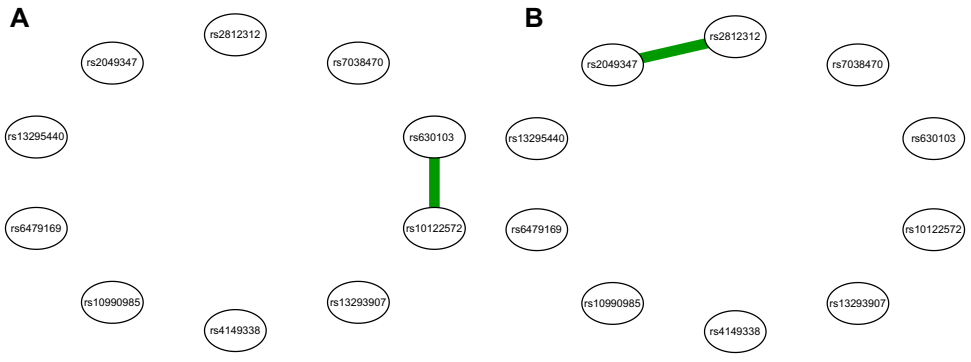

Figure 4. Epistasis network for the phase 1 head and neck cancer GWAS. (**A**) Epistasis network – information theory approach. (**B**) Epistasis network – logistic regression approach.

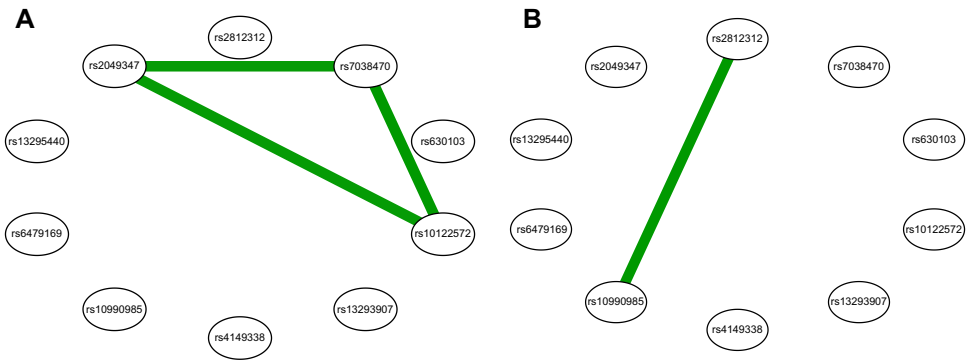

Figure 5. Epistasis network for the phase 2 head and neck cancer GWAS. (**A**) Epistasis network – information theory approach. (**B**) Epistasis network – logistic regression approach.

the presence of only main effects, the interactions that included SNPs without main effects were not identifiable using this approach. In contrast, the logistic regression approach always included the true simulated interactions; however, it also included a higher number of false positives compared to the information theory approach. The higher number of false positives could be due to the fact that the logistic regression was performed using a single interaction at a time instead of including all the interactions in a single multivariable regression model. This would lead to model misspecification in the logistic regression framework. Importantly, covariates can be easily incorporated into the logistic regression approach, whereas inclusion of covariates is not straightforward in the information theory approach. The presence of SNPs in low linkage disequilibrium ($r^2 < 0.1$) had little effect on the overall conclusions. However, when some of the SNPs were in high linkage disequilibrium, the epistasis was confounded with the linkage disequilibrium. Finally, in this work we considered information gain greater than zero to be a significant interaction; however, alternatively, one could evaluate the significance of epistasis by computing the null distribution through permutation of the case–control labels.

We applied the two approaches to develop epistasis networks for the head and neck cancer genetic data collected in two phases. The discrepancies between the logistic regression approach and the information theory approach were due to SNP *rs2812312* having a significant main effect. Therefore, the interactions including SNP *rs2812312* were possibly not identified by the epistasis networks modeled using the information theory-based approach, which is consistent with our observations from the simulation study. Furthermore, the epistasis networks identified using the data from phase 1 were not replicated when we used the data from phase 2. This might have occurred because of the low power to detect epistasis in human GWAS data.[32]

In summary, we have provided insights into the construction of epistasis networks using the information theory approach and the logistic regression approach. We concluded that the information theory approach more efficiently detects interaction effects when main effects are absent. In general, the logistic regression approach is appropriate in all scenarios but results in higher false positives. An understanding of the various strengths and weaknesses of these approaches provides insight for developing novel sophisticated methods to identify epistasis networks.

Acknowledgments

We thank Lee Ann Chastain for editing the manuscript. Some of the controls used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1 through dbGaP accession number phs000092.v1.p. Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI) (U01 HG004422). SAGE is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract "High-throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C).

Author Contributions

Conceived and designed the experiments: RT, SS. Analyzed the data: RT, SS. Wrote the first draft of the manuscript: RT,

SS. Contributed to the writing of the manuscript: RT, SS. Agree with manuscript results and conclusions: RT, SS. Jointly developed the structure and arguments for the paper: RT, SS. Made critical revisions and approved final version: SS. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Easton DF, Eeles RA. Genome-wide association studies in cancer. *Hum Mol Genet*. 2008;17(R2):R109–15.
2. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*. 2002;11(20):2463–8.
3. Phillips PC. The language of gene interaction. *Genetics*. 1998;149(3):1167–71.
4. Wang X, Elston RC, Zhu X. The meaning of interaction. *Hum Hered*. 2010;70(4):269–77.
5. Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. *Ann Med*. 2002;34(2):88–95.
6. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*. 2001;69(1):138–47.
7. Nagel RL. Epistasis and the genetics of human diseases. *C R Biol*. 2005;328(7):606–15.
8. Hemani G, Knott S, Haley C. An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet*. 2013;9(2):e1003295.
9. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109(4):1193–8.
10. Verhoeven KJ, Casella G, McIntyre LM. Epistasis: obstacle or advantage for mapping complex traits? *PLoS One*. 2010;5(8):e12264.
11. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet*. 2003;4(9):701–9.
12. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol*. 2008;32(7):615–26.
13. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008;64(3):685–94.
14. Zhao JY, Jin L, Xiong MM. Test for interaction between two unlinked loci. *Am J Hum Genet*. 2006;79(5):831–845.
15. Chen GK, Thomas DC. Using biological knowledge to discover higher order interactions in genetic association studies. *Genet Epidemiol*. 2010;34(8):863–78.
16. Ferreira T, Marchini J. Modeling interactions with known risk loci-a Bayesian model averaging approach. *Ann Hum Genet*. 2011;75(1):1–9.
17. Li J, Zhang K, Yi N. A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies. *Hum Hered*. 2011;71(3):148–60.
18. Yi N, Kaklamani VG, Pasche B. Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Ann Hum Genet*. 2011;75(1):90–104.
19. Zhang Y. A novel Bayesian graphical model for genome-wide multi-SNP association mapping. *Genet Epidemiol*. 2012;36(1):36–47.
20. Zhang Y, Jiang B, Zhu J, Liu JS. Bayesian models for detecting epistatic interactions from genetic data. *Ann Hum Genet*. 2011;75(1):183–93.
21. Culverhouse RC. A comparison of methods sensitive to interactions with small main effects. *Genet Epidemiol*. 2012;36(4):303–11.
22. Fang G, Haznadar M, Wang W, et al. High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. *PLoS One*. 2012;7(4):e33531.
23. Knights J, Yang J, Chanda P, Zhang A, Ramanathan M. Symphony, an information-theoretic method for gene-gene and gene-environment interaction analysis of disease syndromes. *Heredity*. 2013;110(6):548–59.
24. Molinaro AM, Carriero N, Bjornson R, Hartge P, Rothman N, Chatterjee N. Power of data mining methods to detect genetic associations and interactions. *Hum Hered*. 2011;72(2):85–97.
25. Schwarz DF, Konig IR, Ziegler A. On safari to random jungle: a fast implementation of random forests for high-dimensional data (vol 26, pg 1752, 2010). *Bioinformatics*. 2011;27(3):439.
26. Shervais S, Kramer PL, Westaway SK, Cox NJ, Zwick M. Reconstructability analysis as a tool for identifying gene-gene interactions in studies of human diseases. *Stat Appl Genet Mol Biol*. 2010;9(1):Article18.
27. Zhu Z, Tong X, Zhu Z, et al. Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PLoS One*. 2013;8(4):e61943.
28. John JMM, Van Lishout F, Van Steen K. Model-based multifactor dimensionality reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. *Eur J Hum Genet*. 2011;19(6):696–703.
29. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*. 2011;12:364.
30. Edwards S. Elements of information theory, 2nd edition. *Inf Process Manage*. 2008;44(1):400–1.
31. Sloane BF, Moin K, Krepela E, Rozhin J. Cathepsin-B and its endogenous inhibitors – the role in tumor malignancy. *Cancer Metastasis Rev*. 1990;9(4):333–52.
32. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet*. 2014;15(1):22–33.