Bioinformatics and Biology Insights

SHORT REPORT

Simple F Test Reveals Gene-Gene Interactions in Case-Control Studies

Guanjie Chen[1], Ao Yuan[2], Jie Zhou[1], Amy R. Bentley[1], Adebowale Adeyemo[1] and Charles N. Rotimi[1]

[1]Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, Maryland, USA. [2]National Human Genome Center, Howard University, Washington DC, USA.
Corresponding author email: chengu@mail.nih.gov; rotimic@mail.nih.gov

Abstract: Missing heritability is still a challenge for Genome Wide Association Studies (GWAS). Gene-gene interactions may partially explain this residual genetic influence and contribute broadly to complex disease. To analyze the gene-gene interactions in case-control studies of complex disease, we propose a simple, non-parametric method that utilizes the F-statistic. This approach consists of three steps. First, we examine the joint distribution of a pair of SNPs in cases and controls separately. Second, an F-test is used to evaluate the ratio of dependence in cases to that of controls. Finally, results are adjusted for multiple tests. This method was used to evaluate gene-gene interactions that are associated with risk of Type 2 Diabetes among African Americans in the Howard University Family Study. We identified 18 gene-gene interactions ($P < 0.0001$). Compared with the commonly-used logistical regression method, we demonstrate that the F-ratio test is an efficient approach to measuring gene-gene interactions, especially for studies with limited sample size.

Keywords: F-ratio test, g-g interactions, power

Bioinformatics and Biology Insights 2012:6 169–176

doi: 10.4137/BBI.S9867

This article is available from http://www.la-press.com.

Introduction

Both genetic and environmental risk factors play critical roles in the development of human diseases. Understanding the etiology of complex diseases, such as Type 2 diabetes (T2D), is proving to be a challenging task.[1–4] Partly responsible for this difficulty is the current inability to systematically account for genetic effects that manifest solely or partially in interaction with other genes.[5] Many studies[6–8] suggest that gene-gene interactions may play an important role in disease etiology. As such, the development of statistical tools to detect these genetic effects has received considerable attention. One of the most commonly-used methods for identifying gene-gene interactions is logistic regression, which models the relationship between genotypes and qualitative clinical outcomes.[9–15] Although convenient in application and efficient in inference when the model represents the true relationship in the population, there are a few limitations to this method that should be considered. First, a major challenge of parametric methods, like the logistic model, is the robustness and reliability of the modeling. It is known, for example,[16,17] that when a given model does not represent the true relationships in the population being evaluated, bias will be introduced. This is a serious issue in practice when researchers are not sure of the validity of the underlying parametric model. Model justification, except for very simple cases, is a daunting task, especially with multi-dimensional data. Second, the number of possible interaction terms grows exponentially with the addition of each main effect; logistic regression is limited with regards to interaction data involving many simultaneous factors.[18,19] Third, parametric approaches have less power for detecting interactions than independent main effects, necessitating large sample sizes.[20] Finally, interpreting the parameter estimates for interaction terms resulting from this type of analysis is not straightforward.[21] In contrast, the non-parametric approach, although generally requiring larger sample sizes than parametric methods, are robust and reliable and have been successfully used in genetic analysis. Non-parametric methods are generally more complicated in formulation and computation than parametric methods, due to the non-parametric modeling of the data distribution. However, it is usually simpler to construct the test statistic and compute results for hypothesis testing using non-parametric methods, as accurate asymptotic results can be applied without concern over robustness. Here we present a non-parametric, model-free approach to detect gene-gene interactions in case-control studies. When both case and control SNP frequencies are in Hardy-Weinburg equilibrium (HWE), the test statistic is simplified to a standard F-distribution by asymptotic approximation; when the SNPs are not in HWE, the test statistic approximates a non-centralized F-distribution. The corresponding *P*-value and its power under the alternative can easily be computed via simulation. We demonstrate this method in an analysis of T2D in the Howard University Family Study (HUFS).[22]

The Method

Let SNP1 and SNP2 be trait-related loci, with genotypes represented by values of 0, 1, and 2. Let (x_{11}, x_{21}), ..., (x_{1n}, x_{2n}) be genotypes for SNP1 and SNP2 among cases, while (y_{11}, y_{21}), ..., (y_{1n}, y_{2n}) are the genotypes of SNPs among controls. To investigate whether a SNP by SNP interaction influences the outcome of interest, we will determine whether a joint frequency of these SNPs differs by case status. Let H_0 be the hypothesis that the two SNPs are independent. Statistically, this can be tested by constructing two 3 by 3 contingency tables. For the cases, the (i, j)-cell is the count $n_{ij}^{(1)}$ among the (x_{1k}, x_{2k}) takes value $(i,j)\,(i,j=0,1,2)$, and controls counts $n_{ij}^{(2)}\,(i,j=0,1,2)$. Let $n_i^{(1)}=\sum_{j=0}^{2} n_{ij}^{(1)}\,(i=0,1,2), n_i^{(2)}=\sum_{j=0}^{2} n_{ij}^{(2)}\,(i=0,1,2)$, and

$$\chi_1^2 = n\sum_{i=0}^{2}\frac{\left(n_{ii}^{(1)}-n_i^{2(1)}/n\right)^2}{n_i^{2(1)}} + 2n\sum_{i<j}\frac{\left(n_{ij}^{(1)}-2n_i^{(1)}n_j^{(1)}/n\right)^2}{2n_i^{(1)}n_j^{(1)}},$$

and

$$\chi_2^2 = n\sum_{i=0}^{2}\frac{\left(n_{ii}^{(2)}-n_i^{2(2)}/n\right)^2}{n_i^{2(2)}} + 2n\sum_{i<j}\frac{\left(n_{ij}^{(2)}-2n_i^{(2)}n_j^{(2)}/n\right)^2}{2n_i^{(2)}n_j^{(2)}}.$$

Under H_0, both the case and control cell counts will be in Hardy-Weinberg equilibrium, thus χ_1^2 and χ_2^2 will be asymptotically independent chi-square distribution with degree of freedom three, so asymptotically,

$$T = \chi_1^2/\chi_2^2 \sim F_{3,3}$$

a F distribution with degree (3,3), and if H_0 is true, this statistic will be close to 1; If H_0 is not true, it will deviate significantly from 1. For relevant *P*-value for a specific level of a (typically, $\alpha = 0.05$, 0.03, 0.02 or 0.01), can be determined using an $F_{3,3}$ table.

To quantify the magnitude of the interaction, we may define $r = 2T/(1 + T) - 1$ as a measurement for this. Note $-1 \le r \le 1$, thus, $r = 0$ corresponds to no interaction, $r = -1$ is the maximum negative correlation, and $r = 1$ is the maximum positive correlation.

Note that spurious interactions may occur as a result of SNPs being in linkage disequilibrium (LD) with each other. While LD could first be tested among controls, this step is not necessary with this method. In the absence of an interaction, LD should not differ between cases and controls and, as the test statistic is the ratio of cases to controls, LD should not affect the results.

Deviation from Hardy-Weinberg equilibrium is possible for reasons other then linkage to the trait. In this situation, χ_1^2 will be an asymptotically independent non-central chisquare distribution with 3 degree of freedom, with parameter of non-centrality

$$n\delta_1 = n\sum_{i=0}^{2}\frac{(p_{ii} - p_i^2)^2}{p_i^2} + 2n\sum_{i<j}\frac{(p_{ij} - 2p_ip_j)^2}{2p_ip_j},$$

where p_i is the frequency for SNP i ($i = 0, 1, 2$), and p_{ij} is the frequency of joint SNP type (i, j) ($i, j = 0, 1, 2$) for the cases. Similarly, χ_2^2 will be asymptotically independent non-central chisquare distribution with 3 degree of freedom, with parameter of non-centrality

$$n\delta_2 = n\sum_{i=0}^{2}\frac{(q_{ii} - q_i^2)^2}{q_i^2} + 2n\sum_{i<j}\frac{(q_{ij} - 2q_iq_j)^2}{2q_iq_j},$$

where q_i is the frequency for SNP i ($i = 0, 1, 2$), and q_{ij} is the frequency of joint SNP type (i, j) ($i, j = 0, 1, 2$) for the controls.

So asymptotically,

$$T = \chi_1^2/\chi_2^2 \sim F_{3,3}(n\delta_1, n\delta_2)$$

follows an *F* distribution with degrees of freedom (3,3) and non-centrality parameters $n\delta_1$ and $n\delta_2$. Under H_0, $p_i = q_i, p_{ij} = q_{ij}$ $(i, j = 0, 1, 2)$, so $\delta_1 = \delta_2$, the ratio will be close to 1. If H_0 is not true, typically $\delta_1 > \delta_2$, the ratio will tend to deviate from 1 significantly. For given data, n, and (δ_2, δ_2), the *P*-value of the observed ratio and the power of the level α test can be computed via simulation.

Specifically, under H_0, for each given $\delta_1 = \delta_2 = \delta$, the *P*-value of the observed statistic T is computed as below. Choose a large m (typically, $m = 100, 000$), for $j = 1, ..., m$, do the following:

i. Sample $X_{j,k}$ and $Y_{j,k}$ independently from $N((n\delta/3)^{1/2}, 1)$, $(k = 1, 2, 3)$. Let $Z_j = (X_{j,1}^2 + X_{j,2}^2 + X_{j,3}^2)/(Y_{j,1}^2 + Y_{j,2}^2 + Y_{j,3}^2)$, then Z_j is a sample from $F_{3,3}(n\delta, n\delta)$.
ii. Let $V_j = I(Z_j > T)$, here $I(\cdot)$ is the indicator function, ie, V_j takes value 1 if $Z_j > T$, and 0 otherwise.

Then $P(\delta) = \sum_{j=1}^{m} V_j/m$ is the simulated *P*-value at δ of the observed T. Let $Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(m)}$ be the ordered values of the Z_j's. Let $r = [(1 - \alpha)m]$, the largest integer under $(1 - \alpha)m$, the upper $(1 - \alpha)$-th quantile of the $F_{3,3}(n\delta, n\delta)$ distribution at δ is simulated as $Q(1 - \alpha, \delta) = Z_{(r)}$.

The *P*-value can be tabulated for a list of different δ's, for example, for $\delta = 0.1, 0.2, \ldots$

Similarly, for given $\delta_1 > \delta_2$, the power of the level α test is simulated as below. For $j = 1, \ldots, m$, do the following:

i. Sample $X_{j,k}$ $(k = 1, 2, 3)$ independently from $N((n\delta_1/3)^{1/2}, 1)$, and $Y_{j,k}$ $(k = 1, 2, 3)$ independently from $N((n\delta_2/3)^{1/2}, 1)$. Let $Z_j(X_{j,1}^2 + X_{j,2}^2 + X_{j,3}^2)/(Y_{j,1}^2 + Y_{j,2}^2 + Y_{j,3}^2)$, then Z_j is a sample from $F_{3,3}(n\delta_1, n\delta_2)$.
ii. Let $V_j = I(Z_j > Q(1 - \alpha, \delta_2))$, then $P(\delta_1, \delta_2) = \sum_{j=1}^{m} V_j/m$ is simulated power at (δ_1, δ_2). Here $Q(1 - \alpha, \delta_2)$ is computed before.

For given level of α, let $F(1 - \alpha)$ be the $(1 - \alpha)$-th quantile, the rejection rule for H_0 is

$$T > F(1 - \alpha)$$

and the power $\beta(\delta)$, when the true data is generated with $\delta > 0$, is

$$\beta(\delta) = P_\delta(T > F(1 - \alpha)).$$

The power at a given level of α can be tabulated for a list of different (δ_1, δ_2)'s, and n's for example, for $(\delta_1, \delta_2) = (0.1, 0), (0.2, 0), \ldots, (1, 0)$, and for $n = 30, 50, 100, 150, 200\ldots$

When one (or both) of the minor alleles for the SNP pair being tested has a small frequency, the rare homozygote SNP type will have extremely small frequency in the contingency table. In this case, the asymptotic approximation of the F-distribution for the T statistic is not justified. Let n_0 be the smallest observed frequency in either the case and control contingency Tables. As a rule of thumb, when $n_0 < 10$, the sample size is not large enough for the asymptotic approximation to be valid. In this case, the 'exact' P-value (under the null) of the observed statistic T can be computed by the standard exact method.

Departures from Hardy-Weinberg equilibrium among controls was assessed by comparing the observed genotype frequencies to the expected frequencies using the exact test. Odds ratio and 95% confidence intervals for single locus associations were obtained using unconditional logistic regression. As a basis for comparison, logistic regression models were also performed to evaluate the gene-gene interactions. Models included each SNP individually as well as a SNP × SNP product term.

The FDR method was used to adjust for multiple testing,[23] although, if all the tests are independent, a Bonferroni correction may also used.[24]

Analysis and the software used are written in SAS and can be provided upon request to chengu@mail.nih.gov.

Data Analysis

We applied our method to T2D using the Howard University Family Study (HUFS) data.[22] Briefly, the HUFS is a population based family study of African Americans in the Washington, D.C. metropolitan area. The major objective of the HUFS was to enroll and examine a randomly-ascertained sample of African American families, along with a set of unrelated individuals, for the study of the genetic and environmental bases of common complex diseases including hypertension, obesity, diabetes and associated phenotypes. A total of 1082 unrelated individuals had both phenotype and genotype (Affymetrix 6.0) data. Of these, 221 individuals were classified as T2D (defined as fasting plasma glucose concentration > 126 mg/dL, report of a doctor's diagnosis of T2D, or report of current T2D treatment).

Based on previous publications,[25,26] 19 T2D candidate gene regions (Table 1) were selected for analysis. Of note, the issue of loci interaction is independent from consideration of main effect: loci that strongly interact may or may not be associated individually with the trait. Thus, the SNPs included in our analysis were not first limited to those with a main effect on T2D. Of these, 608 SNPs passed quality control filters: call rate ≥ 95%, Minor Allele Frequency (MAF > 0.05), and Hardy-Weinberg Equilibrium (P-values of HWE > 0.01). After using window size

Table 1. The list of candidate genes that were analyzed.

Genes	Location	No. of SNPs	Order*
GCKR	2p23	4	1–4
BCL11A	2p16.1	9	5–13
IRS1	2q36	6	14–19
PPARG	3p25	15	20–34
WFS1	4p16.1	8	35–42
KLF14	7q32.3	1	43–43
TP53INP1	8q22	3	44–46
TCF7L2	10q25.3	30	47–76
KCNQ1	11p15.5	71	77–147
KCNJ11	11p15.1	3	148–150
CENTD2/ARAP1	11q13.4	3	151–153
MTNR1B	11q21	4	154–157
HMGA2	12q15	15	158–172
IGF1	12q23.2	8	173–180
HNF1A	12q24.2	3	181–183
ZFAND6	15q25.1	7	184–190
PRC1	15q26.1	6	191–196
FTO	16q12.2	78	197–274
HNF1B	17q21.3	24	275–298

Note: *The order represents the position of the SNP in Figures 1 and 2.

Table 2. Significant results for single locus association of 298 SNPs in 19 genes.

SNPs	Genes	Odds ratio	95% C.I.	*P*-values
rs10956932	*TP53INP1*	1.62	1.27–2.05	0.00008
rs8053888	*FTO*	0.66	0.11–0.53	0.00025
rs12573128	*TCF7L2*	0.67	0.53–0.84	0.00070
rs231901	*KCNQ1*	0.49	0.21–0.75	0.00092
rs9806929	*FTO*	0.60	0.42–0.84	0.00284
rs11649763	*KNF1B*	0.44	0.25–0.77	0.00403
rs7069007	*TCF7L2*	1.56	1.13–2.15	0.00738
rs5742652	*IGF1*	2.08	1.19–3.62	0.00981

Table 3. Top results of gene-gene interactions from 298 SNPs in 19 genes.

Locus (name of genes)	Locus (name of genes)	*P*-value
rs10519280 (*ZFAND6*)	rs12149010 (*FTO*)	2.62×10^{-6}
rs5742652 (*IGF1*)	rs7205617 (*FTO*)	7.71×10^{-6}
rs17130192 (*TCF7L2*)	rs12425829 (*HMGA2*)	8.27×10^{-6}
rs17130192 (*TCF7L2*)	rs11111262 (*IGF1*)	8.57×10^{-6}
rs17130192 (*TCF7L2*)	rs17636091 (*PRC1*)	1.10×10^{-5}
rs2272046 (*HMGA2*)	rs17636091 (*PRC1*)	1.35×10^{-5}
rs2272046 (*HMGA2*)	rs6824720 (*WFS1*)	1.45×10^{-5}

of 50 and R^2 score ≥ 0.3 between two loci, 298 SNPs not in LD with each other were used for analysis (Table 1) in 19 candidate T2D gene regions.

For reference, logistic regression analysis of each of the loci without interaction was conducted (all results $P < 0.01$ are presented in Table 2). After correction for multiple tests, no SNP reached the threshold for statistical significance (Bonferroni significant level $P < 1.7 \times 10^{-4}$). The threshold for statistical significance for the gene-gene interaction evaluated by the F-ratio method was set at $P < 0.001$ (corresponding to an FDR q-value of 0.027); at this level of statistical significance, the dependence between the two loci among cases was over 141 times higher than among controls. 18 significant gene-gene interactions were discovered (the top 7 are presented in Table 3). For comparison, logistic regression was also used to evaluate gene-gene interactions in the same data. To illutrate the overall similarity of these approaches, a heat map was created showing the statistical significance of the interaction term for each pair of SNPs evaluated using the F-ratio (Fig. 1) and logistic regression (Fig. 2) analyses. Similar patterns were observed with both of these methods; at the same level of statistical significance ($P = 0.05$), there was a concordance rate of 94.09% between the two methods. The generally lower *P*-values observed with logistic regression are presumed to represent the fact that logistic regression models are already adjusted for the main effect of each of the SNPs, while the F-ratio method is not. Displayed in Figure 3 is the power of the F-ratio method for a variety of δ values (a measure of the deviation from HWE between two SNPs), sample size, and α levels. At an $\alpha = 0.05$,

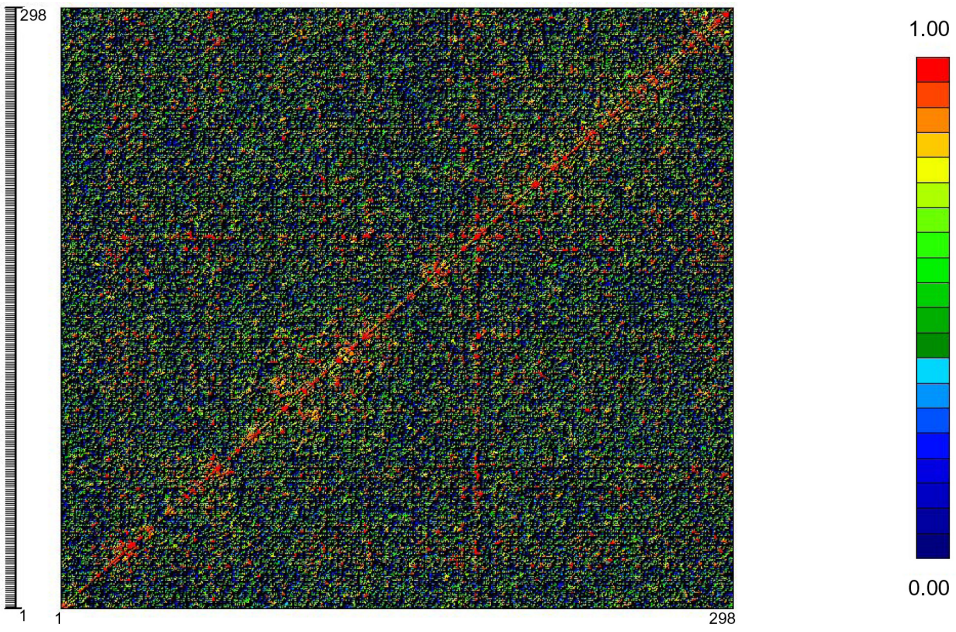

Figure 1. Gene-gene interactions among 298 SNPs distributed in 19 genes using our simple F-ratio non-parametric method.
Note: Colors from dark blue to red represent *P*-value from 0.00 to 1.00.

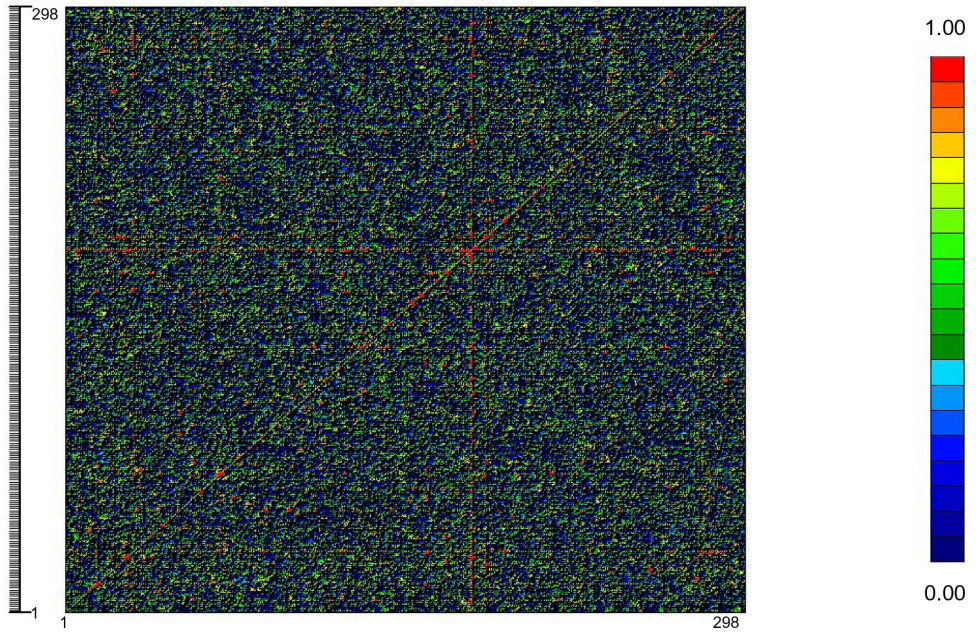

Figure 2. Gene-gene interactions among 298 SNPs which distributed on 19 genes using logistic regression.
Note: Colors from dark blue to red represent *P*-value from 0.00 to 1.00.

a $\delta_1 = 0.4$, and a sample size of $n = 100$, the F-ratio method reaches 0.80 power. The strong power that can be achieved at this moderate value of δ with less than 200 individuals suggests the practicality of using this method when sample size is limited.

Discussion

We present a new method for evaluating gene-gene interactions that uses the F-ratio test. Using this method, 18 gene-gene interactions were found to influence risk of Type 2 diabetes among African

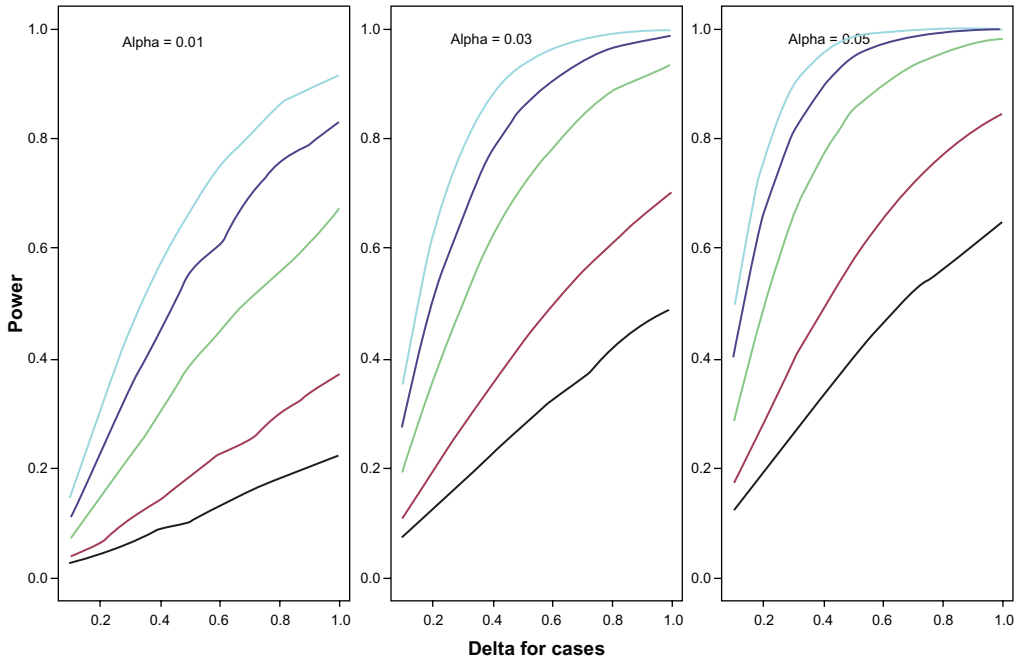

Figure 3. Powers of simple F-ratio test.
Notes: x-axis: δ values in cases ($\delta = 0$ in controls). *y*-axis: the powers. Panels from left to right represent α of 0.01, 0.03 and 0.05. The color of the lines indicate sample size: black (n = 30), red (n = 50), green (n = 100), purple (n = 150), and blue (n = 200).

Americans of the Howard University Family Study. As each of the genes investigated are candidate genes, their individual role in disease risk is presumed. Identifying the specific mechanisms by which these genes would be expected to interact is beyond the scope of this work, but the top results suggest that some of the effect of genes involving in insulin sensitivity (such as *ZFAND*6, and *IGF*1) is mediated through obesity (*FTO*)[26,27] a reasonable hypothesis. In comparison with logistic regression, the F-ratio test was shown to be an efficient method with minimal potential bias and good power to detect moderate gene-gene interaction even in relatively small sample sizes.

An exhaustive investigation of all pairwise loci interactions search in genome-wide data is time consuming. Given 500,000 to 1,000,000 SNPs in 5,000 individuals, computation time may be several weeks or even months.[21] Although the F-ratio method does not decrease the number of tests, it significantly reduces CPU time per test from 0.04 (logistic regression) to 0.01 (F-ratio method) seconds in the same computing environments.

The results of gene-gene interaction analysis were corrected by using the FDR method. As SNPs in LD were excluded from the analysis in order to increase effciency, a Bonferroni correction could have been used[28] [correcting for $[(\#\text{ of locus})^2 - (\#\text{ of locus})]/2$ tests]. Using Bonferroni correction would be overly conservative; the existance of marginal effects negates the multiple testing cost.[24]

Conclusion

The F-ratio test was used as a non-parametric method for comparing the relationship between trait-associated loci in cases to that in controls. A different pattern of joint genotype frequencies in cases compared to controls indicates an interaction between these loci that affects case status. This method represents a novel technique to identify the combination of polymorphisms associated with the risk of common complex diseases. This method overcomes some limitations of logistic regression modeling for detection and characterization of gene-gene interactions. The F-ratio method performed well in Type 2 Diabetes case-control data, identifying 18 gene-gene interactions. This F-ratio test is a useful statistical tool for the analysis of gene-gene interactions and represents a significant contribution in the context of the heritability that remains unexplained by single locus association studies.

Acknowledgement

The study was supported by grants S06GM008016-320107 to CNR and S06GM008016-380111 to AA, both from the NIGMS/MBRS/SCORE Program. Participant enrollment was carried out at the Howard University General Clinical Research Center (GCRC) which is supported by grant number 2M01RR010284 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the offcial view of NCRR or NIH. Additional support was provided by the Coriell Institute for Medical Research. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, in the Center for Research in Genomics and Global Health.

Author Contributions

Conceived and designed the experiments: GC, AY, AA, CR. Analysed the data: GC, JZ. Wrote the first draft of the manuscript: GC. Contributed to the writing of the manuscript: AB, AY, AA, CR. Agree with manuscript results and conclusions: All authors. Jointly developed the structure and arguments for the paper: GC, AY, AB. Made critical revisions and approved final version: AY, AB, AA, CR. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors

have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet*. 2001;69:936–50.
2. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. 2002;4:45–61.
3. Joannidis JP, Trikalinos TA, Ntzani EE. Contopouloc-Ioannidis DG, Genetic association in large versus small studies: an empirical assessment. *Lancet*. 2003;361:567–71.
4. Wang WYS, Barratt BJ, Todd JA. Genome-Wide association studies: theoretical issues and practical concerns. *Nat Rev Genet*. 2003:109–18.
5. Templeton AR. Epistasis and complex traits. In: Wade M, Brodie B III, Wolf J, editors. *Epistasis and Evolutionary Process*. Oxford University Press, Oxford; 2000:41–57.
6. Jarvik G, Larson EB, Goddard K, Schellenberg GD, Wijsman EM. Influence of apolipoprotein E genotype on the transmission of Alzheimer disease in a community-based sample. *Am J Hum Genet*. 1996;58:191–200.
7. MacLeod S, Sinha R, Kadlubar FF, Lang NP. Polymorphism of CTP1A1 and GSTM1 influence the in vivo function of CYP1A2. *Mutat Res*. 1997;376:135–42.
8. Beales PL, Kopelman PG. Obesity genes. *Clin Endocrinol*. 1996;45:373–8.
9. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons, Inc; 2000.
10. Holmans P. Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity*. 2002;53:92–102.
11. Kang JH, Kim MJ, Ko SH, et al. Upregulation of rat Ccnd1 gene by exendim-4 in pancreatic beta cell line INS-1: interaction of early growth response-1 with cis-regulatory element. *Diabetologia.* March 18, 2006.
12. Huang WY, Berndt SI, Kang D, et al. Nucleotide excision repair gene polymorphisms and risk of advanced colorectal adenoma: *XPC* polymorphisms modify smoking-related risk. *Cancer Epidemiology Biomarkers and Prevention*. 2006;15:306–11.
13. Li MC, Cui ZS, He QC, Zhou BS. Association of genetic polymorphism in the DNA repair gene XRCC1 with susceptivility to lung cancer in non-smoking women. *Zhonghua Zhong Liu Za Shi*. 2005;12(12):713–6.
14. Wang C, Zhou X, Ye S, et al. Combined effects of apo E-CI-CII cluster and LDL-R gene polymorphisms on chromosome 19 and coronary artery disease risk. *Int J Hyg Environ Health*. February 2, 2006.
15. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different bilogical data and computaional classification methods for use in protein interaction prediction. *Proteins*. January 31, 2006.
16. Huber P. The behavior of maximum likelihood estimates under nonstandard conditions. *Proc Fifth Berkeley Symp Math Statist Probab*. 1967;1:221–33.
17. Pfanzagl J. On the measurability and consistency of minimum contrast estimators. *Metrika*. 1969;14:249–72.
18. Rttchie DM. Bioinformatics approaches for detecting gene-gene and gene-enviroment interactions in studies of human disease. *Neurosurg Focus*. 2005;19(4):E2.
19. Moore HJ. Computational approaches to detecting and characterizing gene-gene interactions. PSB 2003 Tutotial.
20. Moore HJ, Gilbert CJ, Tsai CT, et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol*. January 31, 2006.
21. Cordell JH. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009;10(6):392–404.
22. Adeyemo A, Gerry N, Chen G, et al. A Genome Wide association Study of hypertension and blood pressure in African Americans. *PLoS Genet*. 2009;5(7):e1000564.
23. Storey J, Tibshirani R. Statistical significance for genome wide studies. *Proc Natl Acad Sci*. 2003;100:9440–5
24. Marchini J, Donnelly P, Cardon LR. Genome wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet*. 2005;37:413–7.
25. Parikh H, Groop L. Candidate genes for type 2 diabetes. *Reviews in Endocrine and Metabolic Disorders*. 2004;5:151–76.
26. Voight B, Scott L, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large scale association analysis. *Nat Genet*. 2010;42(7):579–89.
27. Clemmons DR. Role of insulin-like growth factor in maintaining normal glucose homeostasis. *Horm Res*. 2004;62(Suppl 1):77–82.
28. Abdi H. Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ, editor. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage; 2007.