Epistasis analysis for quantitative traits by functional regression model

Futao Zhang,[1,2] Eric Boerwinkle,[2] and Momiao Xiong[2,3]

[1]Institute of Bioinformatics, Zhejiang University, Hangzhou, Zhejiang 310058, China; [2]Human Genetics Center, Division of Biostatistics, The University of Texas School of Public Health, Houston, Texas 77030, USA

The critical barrier in interaction analysis for rare variants is that most traditional statistical methods for testing interactions were originally designed for testing the interaction between common variants and are difficult to apply to rare variants because of their prohibitive computational time and poor ability. The great challenges for successful detection of interactions with next-generation sequencing (NGS) data are (1) lack of methods for interaction analysis with rare variants, (2) severe multiple testing, and (3) time-consuming computations. To meet these challenges, we shift the paradigm of interaction analysis between two loci to interaction analysis between two sets of loci or genomic regions and collectively test interactions between all possible pairs of SNPs within two genomic regions. In other words, we take a genome region as a basic unit of interaction analysis and use high-dimensional data reduction and functional data analysis techniques to develop a novel functional regression model to collectively test interactions between all possible pairs of single nucleotide polymorphisms (SNPs) within two genome regions. By intensive simulations, we demonstrate that the functional regression models for interaction analysis of the quantitative trait have the correct type 1 error rates and a much better ability to detect interactions than the current pairwise interaction analysis. The proposed method was applied to exome sequence data from the NHLBI's Exome Sequencing Project (ESP) and CHARGE-S study. We discovered 27 pairs of genes showing significant interactions after applying the Bonferroni correction (P-values $< 4.58 \times 10^{-10}$) in the ESP, and 11 were replicated in the CHARGE-S study.

[Supplemental material is available for this article.]

Epistasis is the primary factor in molecular evolution (Breen et al. 2012) and plays an important role in quantitative genetic analysis (Steen 2011). Epistasis is a phenomenon in which the effect of one genetic variant is masked or modified by one or more genetic variants and is often defined as the departure from additive effects in a linear model (Fisher 1918). Many statistical methods, including regression-based methods, have been developed to detect epistasis in quantitative genetic analysis (Cordell 2009; Chen and Cui 2010; Bocianowski 2012). However, these methods were originally designed to detect epistasis for common variants (Steen 2011) and are difficult to apply to rare variants because of their high type 1 error rates and poor ability to detect interactions between rare variants. Next-generation sequencing (NGS) data raise two serious problems. The first problem is the curse of dimensionality of the data, and the second problem comes from the low frequencies of rare variants in the data.

The recently reported average number of single nucleotide polymorphisms (SNPs) per kb in the 202 drug target genes sequenced in 12,514 European subjects is about 48 SNPs (Nelson et al. 2012). The total number of all possible pairs of SNPs across the genome for large sample sizes can reach as many as 10^{16}. The dimension of whole-genome sequencing is extremely high. The high dimension of the data for interaction analysis poses two great challenges. The first challenge is the requirement of a prohibitive amount of computational time. Suppose that 5000 pairwise tests can be finished in 1 sec (Steen 2011), then a possible pairwise interaction test would take ~65,956 yr to finish. The second challenge for genome-wide interaction analysis arises from the multi-

ple statistical tests. The power of the statistics that exhaustively test all possible pairs of interaction will be severely hampered by extremely large numbers of multiple tests.

The popular strategies for reducing the dimensionality of the data, the number of tests, and the time of computations, and for improving the power to detect interactions are feature extraction, which projects the original high-dimensional data to low-dimensional space (Guyon et al. 2006; Li et al. 2009); feature selection, which selects subsets of variables of interests (Guyon and Elisseeff 2003; Saeys et al. 2007); and possibly, approximately complete testing to reduce computational time (Prabhu and Pe'er 2012).

Feature extraction in association studies with NGS data is often carried out by collapsing multiple variants into a single variable (Li and Leal 2008; Bansal et al. 2010; Luo et al. 2011; Wu et al. 2011). However, much important interaction information may be lost after the multiple variants are collapsed. The collapsing methods may lack the power to detect interactions between variants.

To address the critical barrier in detection of gene–gene interactions with NGS data, we take a genome region (or gene) as a basic unit of interaction analysis and use all the information that can be accessed to collectively test interactions between all possible pairs of SNPs within two genome regions (or genes). This will shift the paradigm of interaction studies from pairwise interaction analysis to region–region (gene–gene) interaction analysis, in which we collectively test interactions between two sets of loci within genomic regions or genes. To effectively reduce the dimension of the

data, unlike the recently proposed group association tests (Li and Leal 2008; Madsen and Browning 2009), which ignore differences in genetic effects between SNPs in different locations, we use genetic variant profiles that will recognize information contained in the physical location of the SNP as a major data form. The densely distributed genetic variants across the genomes in large samples can be viewed as realizations of a Poisson process (Joyce and Tavaré 1995). The densely typed genetic variants in a genomic region for each individual are so close that these genetic variant profiles can be treated as observed data taken from curves (Luo et al. 2012). The genetic variant profiles are called functional. Since standard multivariate statistical analyses often fail with functional data (Ferraty and Romain 2010), we formulate a test for the interaction between two genomic regions in quantitative trait analysis as a functional regression (FRG) model (Ramsay and Silverman 2005) with scalar response. In the FRG model, the genotype functions (genetic variant profiles) are defined as a function of the genomic position of the genetic variants rather than a set of discrete genotype values, and the quantitative trait is predicted by genotype functions with their interaction terms. We will show that the FRG with scale response is a natural extension of the multivariate regression for interaction analysis.

To evaluate its performance for interaction analysis, we use large-scale simulations to calculate the type I error rates of the FRG for testing the interaction between two genomic regions and to compare its power with pairwise interaction analysis and regression on principal components (PCs). To further evaluate its performance, the FRG for interaction analysis is applied to high-density lipoprotein (HDL) and exome sequence data from the NHLBI's Exome Sequencing Project (ESP) and to whole-genome sequencing data from the CHARGE-S project.

Methods

Functional regression model for interaction analysis with a quantitative trait

Consider the two genomic regions $[a_1, b_1]$ and $[a_2, b_2]$. Let y_i be the phenotypic value of a quantitative trait measured on the ith individual. Let t and s be a genomic position in the first and second genomic regions, respectively. Let $x_i(t)$ and $x_i(s)$ be genotype functions of the ith individual in the regions $[a_1, b_1]$ and $[a_2, b_2]$, respectively. The genotype function $x_i(t)$ of the ith individual is defined as

$$X_i(t) = \begin{cases} 0, & mm \\ 1, & Mm \\ 2, & MM, \end{cases}$$

where M and m are two alleles of the SNP at the genomic position t. Recall that a regression model for interaction analysis is defined as

$$y_i = \mu + \sum_{j=1}^{k_1} x_{ij}\alpha_j + \sum_{l=1}^{k_2} z_{il}\beta_l + \sum_{j=1}^{k_1}\sum_{l=1}^{k_2} x_{ij}z_{il}\gamma_{jl} + \varepsilon_i, \quad (1)$$

where μ is an overall mean; α_j is the main genetic additive effect of the jth SNP in the first genomic region; β_l is the main genetic additive effect of the lth SNP in the second genomic region; γ_{jl} is an additive × additive interaction effect between the jth SNP in the first genomic region and the jth SNP in the second genomic region; x_{ij} and z_{il} are indicator variables for the genotypes at the jth SNP and the lth SNP, respectively; and ε_i are independent and identically distributed normal variables with mean of zero and variance

σ^2. Similar to the multiple regression models for interaction analysis with a quantitative trait, the FRG model for a quantitative trait can be defined as

$$y_i = \alpha_0 + \int_T \alpha(t)x_i(t)dt + \int_S \beta(s)x_i(s)ds + \int_T\int_S \gamma(t,s)x_i(t)x_i(s)dtds + \varepsilon_i,$$
$$(2)$$

where α_0 is an overall mean; $\alpha(t)$ and $\beta(s)$ are genetic additive effects of two putative QTLs located at the genomic positions t and s, respectively; $\gamma(t,s)$ is the interaction effect between two putative QTLs located at the genomic positions t and s; $x_i(t)$ and $x_i(s)$ are genotype function; and ε_i are independent and identically distributed normal variables with mean of zero and variance σ^2.

In theory, the genetic additive effect and interaction effect functions can be obtained by variation of theory (Supplemental Note 1). The classical concept of genetic additive variance and interaction variance can be extended to the functional model (Supplemental Note 2). Below we take a numerical approach to estimate the genetic additive and interaction effect functions.

Estimation of interaction effects

We assume that both phenotypes and genotype functions are centered. The genotype functions $x_i(t)$ and $x_i(s)$ are expanded in terms of the orthonormal basis functions as

$$x_i(t) = \sum_{j=1}^{\infty} \xi_{ij}\varphi_j(t) \text{ and } x_i(s) = \sum_{k=1}^{\infty} \eta_{ik}\psi_k(s), \quad (3)$$

where $\varphi_j(t)$ and $\psi_k(s)$ are sequences of the orthonormal basis functions. The expansion coefficients ξ_{ij} and η_{ik} are estimated by

$$\xi_{ij} = \int_T x_i(t)\varphi_j(t)dt \text{ and } \eta_{ik} = \int_S x_i(s)\psi_k(s)ds. \quad (4)$$

In practice, numerical methods for the integral will be used to calculate the expansion coefficients. Substituting Equation 3 into Equation 2, we obtain

$$y_i = \sum_{j=1}^{\infty} \xi_{ij}\alpha_j + \sum_{k=1}^{\infty} \eta_{ik}\beta_k + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \xi_{ij}\eta_{ik}\gamma_{jk} + \varepsilon_i, \quad (5)$$

where $\alpha_j = \int \alpha(t)\varphi_j(t)dt$, $\beta_k = \int \beta(s)\psi_k(s)ds$ and $\gamma_{jk} = \int_T\int_S \gamma(t,s)\varphi_j(t)\psi_k(s)dtds$. The parameters α_j, β_k, and γ_{jk} are referred to as genetic additive and as additive × additive interaction effect scores. These scores can also be viewed as the expansion coefficients of the genetic effect functions with respect to orthonormal basis functions:

$$\alpha(t) = \sum_j \alpha_j\varphi_j(t), \beta(s) = \sum_k \beta_k\psi_k(s) \quad \text{and} \quad \gamma(s,t) = \sum_j\sum_k \gamma_{jk}\varphi_j(s)\psi_k(t).$$
$$(6)$$

Let $Y = [y_1, ..., y_n]^T, \alpha = [\alpha_1, ..., \alpha_J]^T, \beta = [\beta_1, ..., \beta_K]^T, \gamma = [\gamma_{11}, \gamma_{12}..., \gamma_{JK}]^T, b = [\alpha^T, \beta^T, \gamma^T]^T$,

$$\xi = \begin{bmatrix} \xi_{11} & \cdots & \xi_{1J} \\ \cdots & \cdots & \cdots \\ \xi_{n1} & \cdots & \xi_{nJ} \end{bmatrix}, \eta = \begin{bmatrix} \eta_{11} & \cdots & \eta_{1k} \\ \cdots & \cdots & \cdots \\ \eta_{n1} & \cdots & \eta_{nk} \end{bmatrix} \text{ and}$$

$$\Gamma = \begin{bmatrix} \xi_{11}\eta_{11} & \cdots & \xi_{1J}\eta_{1K} \\ \cdots & \cdots & \cdots \\ \xi_{n1}\eta_{n1} & \cdots & \xi_{nJ}\eta_{nK} \end{bmatrix},$$

where the values J and K are chosen such that genotype function expansions can account for 80% of total genetic variation in the first and second genes, respectively.

If we use the above notations, Equation 5 can be reduced to

$$Y = \xi\alpha + \eta\beta + \Gamma\gamma + \varepsilon = Wb + \varepsilon, \tag{7}$$

where $W = [\xi \quad \eta \quad \Gamma]$ and $\varepsilon = [\varepsilon_1, ..., \varepsilon_n]^T$.

Therefore, the interaction models with integrals are transformed to the traditional multivariate regression models (Equation 7) for interaction analysis. The standard least square estimator of b is given by

$$\hat{b} = (W^T W)^{-1} W^T (Y - \bar{Y}), \tag{8}$$

and its variance is

$$Var(\hat{b}) = \hat{\sigma}^2 (W^T W)^{-1}, \tag{9}$$

where

$$\hat{\sigma}^2 = \frac{1}{n - J - K - JK}(Y - \bar{Y})^T[I - W(W^T W)^{-1} W^T](Y - \bar{Y}).$$

Substituting the estimated genetic effect scores $\hat{\alpha}_j$, $\hat{\beta}_k$, and $\hat{\gamma}_{jk}$ into Equation 6 yields the estimated genetic additive effect and additive $\times$ additive interaction effect functions $\hat{\alpha}(t)$, $\hat{\beta}(s)$, and $\hat{\gamma}(t, s)$. If basis functions for expansion of genotype functions are functional PCs or eigenfunctions (Ash and Gardner 1975), then we can estimate the genetic additive and additive $\times$ additive interaction variances in Equation 2 (Supplemental Note 3).

Test statistics

An essential problem in genetic interaction studies of the quantitative trait is to test the interaction between two genomic regions (or genes). Formally, we investigate the problem of testing the following hypothesis:

$$\gamma(t, s) = 0, \forall t \in [a_1, b_1], s \in [a_2, b_2],$$

which is equivalent to testing the hypothesis

$$\gamma = 0,$$

where γ is defined in Equation 7.

Let Λ be the matrix corresponding to the parameter γ of the variance matrix $Var(\hat{b})$ in Equation 9. Define the test statistic for testing the interaction between the two genomic regions $[a_1, b_1]$ and $[a_2, b_2]$ as

$$T_I = \hat{\gamma}^T \Lambda^{-1} \hat{\gamma}. \tag{10}$$

Then, under the null hypothesis $H_0 : \gamma = 0$, T_I is asymptotically distributed as a central $\chi^2_{(JK)}$ distribution if JK components are taken in the expansion Equation 6.

Results

Null distribution of test statistics

In the previous section, we showed that the test statistics T_I are asymptotically distributed as a central $\chi^2_{(JK)}$ distribution. To examine the validity of this statement, we performed a series of simulation studies to compare their empirical levels with the nominal ones.

The type I error rates for rare variants and both common and rare variants were calculated. We assumed the three models: model 1 (without marginal effects), model 2 (with marginal effect of one gene), and model 3 (with marginal effects of two genes to generate a phenotype) (Supplemental Note 4).

We generated a population with 1 million individuals by resampling from 3212 individuals with variants in eight genes selected from the NHLBI's ESP, where the description of eight genes is summarized in Supplemental Table S1. To examine whether presence of the linkage disequilibrium (LD) between SNPs will seriously affect the type 1 error rates, we included some genes with linked variants. The number of sampled individuals range from 500 to 5000, and 5000 simulations were repeated. Table 1 and Supplemental Table S2 summarize the average type I error rates of the test statistics for testing the interaction between two genes with rare variants and mixed common and rare variants over all possible pairs of eight genes (28 pairs of genes), respectively, at the nominal levels $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.001$. These tables show that the type I error rates of the test statistics for testing interactions between two genes with or without marginal effects are not appreciably different from the nominal α levels.

To study the impact of the LD between SNPs, we present Supplemental Table S3. Supplemental Table S3 summarizes the type 1 error rates of the FRG for testing interactions between genes: GBP3 and KANK4. The LD map of genes GBP3 and KANK4 is shown in Supplemental Figure 1. Supplemental Table S2 demonstrates that the presence of LD between genes being tested did not have a significant impact on the type 1 error rates. The impact of the lengths of the genes and sequencing error (on the type 1 error rates) will be limited (Supplemental Note 3).

Power evaluation

To evaluate the performance of the FRG for testing the interaction between two genes or genomic regions for a quantitative trait, simulated data were used to estimate their power to detect a true interaction. A true quantitative genetic model is given as follows. Consider H pairs of quantitative trait loci (QTLs) from two genes (genomic regions). Let Q_{h_1} and q_{h_1} be two alleles at the first QTL,

Table 1. Average type 1 error rates of the statistics for testing interaction between two genes with rare variants

| Model	Sample size	0.05	0.01	0.001
Model 1	500	0.0506	0.0108	0.0011
	1000	0.0491	0.0101	0.0010
	2000	0.0474	0.0096	0.0010
	3000	0.0480	0.0099	0.0008
	4000	0.0494	0.0092	0.0009
	5000	0.0489	0.0095	0.0012
Model 2	500	0.0530	0.0110	0.0012
	1000	0.0489	0.0114	0.0015
	2000	0.0491	0.0098	0.0010
	3000	0.0488	0.0099	0.0009
	4000	0.0482	0.0094	0.0009
	5000	0.0507	0.0100	0.0009
Model 3	500	0.0496	0.0105	0.0012
	1000	0.0471	0.0094	0.0010
	2000	0.0473	0.0094	0.0011
	3000	0.0481	0.0097	0.0009
	4000	0.0468	0.0098	0.0009
	5000	0.0486	0.0093	0.0008
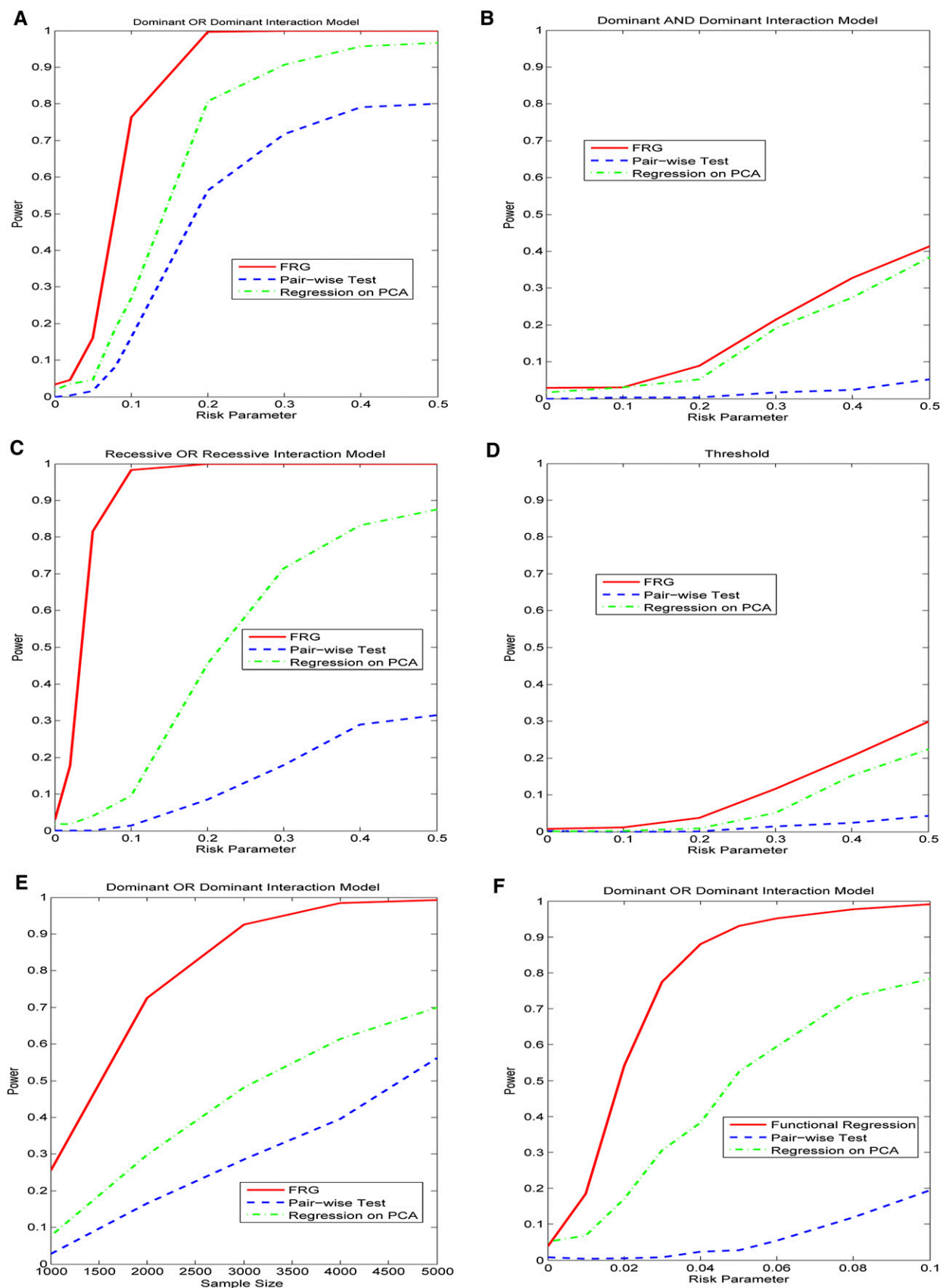
Figure 1. (Legend on next page)

and Q_{h_2} and q_{h_2} be two alleles at the second QTL, for the H pair of QTLs. Let u_{ijkl} be the genotypes of the uth individual with $ij = Q_{h_1}Q_{h_1}, Q_{h_1}q_{h_1}, q_{h_1}q_{h_1}$ and $kl = Q_{h_2}Q_{h_2}, Q_{h_2}q_{h_2}, q_{h_2}q_{h_2}$, and $g_{u_{ijkl}}$ be its genotypic value. The following multiple linear regression is used as a genetic model for a quantitative trait:

$$y_u = \sum_{h=1}^{H} g_{u_{ijkl}}^h + \varepsilon_u, u = 1, 2, ..., n,$$

where $g_{u_{ijkl}}^h$ is a genotypic value of the hth pair of QTLs, and ε_u is distributed as a standard normal distribution $N(0, 1)$.

Four models of interactions are considered: (1) Dominant OR Dominant, (2) Dominant AND Dominant, (3) Recessive OR Recessive, and (4) Threshold model (Supplemental Table S4). The Recessive AND Recessive model is excluded due to infrequency of that condition with rare variants. The parameter r varies from zero to one.

We generated 1 million individuals by resampling from 3212 individuals of European origin with variants in the two genes *IQGAP3* and *ACTN2* selected from the ESP data set. We randomly selected 20% of the variants as causal variants. A total of 2000 individuals for the four interaction models were sampled from the populations. A total of 1000 simulations were repeated for the power calculation.

The power of the proposed method is compared with the regression on PCs. For SNP genotypes in each genomic region, PC analysis (PCA) was performed. The number of PCs for each individual that can explain 80% of the total genetic variation in the genomic region will be selected as the variable. Specifically, the PC score of the ith individual in the first and second genomic regions is denoted by $x_{i1}, ..., x_{ik_1}$ and $z_{i1}, ...z_{ik_2}$, respectively. The regression model for detection of interaction is then given by

$$y_i = \mu + \sum_{j=1}^{k_1} x_{ij}\alpha_j + \sum_{l=1}^{k_2} z_{il}\beta_l + \sum_{j=1}^{k_1}\sum_{l=1}^{k_2} x_{ij}z_{il}\gamma_{jl} + \varepsilon_i.$$

The power of the proposed method is compared with the traditional point-wise interaction test, which takes the following model:

$$y_i = \mu + x_{i1}\alpha_1 + x_{i2}\alpha_2 + x_{i1}x_{i2}\gamma + \varepsilon_i, i = 1, ..., n.$$

For a pair of genes, we assume that the first gene has k_1 SNPs and the second gene has k_2 SNPs, and then the total number of all possible pairs is $k = k_1 \times k_2$. For each pair of SNPs, we calculate a statistic for testing pairwise interaction T_{jpair}. Finally, the maximum of T_{jpair}: $T_{max} = \max(T_{1pair}, T_{2pair}, ..., T_{kpair})$ is computed.

By permutation of 1000 times of the phenotypic values y_i, we can find the distribution of T_{max}; i.e., we have 1000 values of T_{max}. From this empirical distribution, we can find the P-value of T_{max}, which can be used to calculate the power of testing for interaction between two genes (genomic regions) by pairwise tests.

We first study the power of statistics for testing interactions between two genomic regions with all rare variants where 20% of the rare variants were chosen as causal variants. Figure 1, A through D, plots the power curves of three statistics: FRG, regression on PCs, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing interactions between two genomic regions (genes) that consist of only rare variants for a quantitative trait under Dominant OR Dominant, Dominant AND Dominant, Recessive OR Recessive, and Threshold models, respectively. These power curves are a function of the risk parameter at the significance level $\alpha = 0.05$. From these figures we observed several remarkable features. First, under all four interaction models, the test based on the FRG model was the most effective, followed by the regression on PCA. The pairwise tests where we tested the interaction between all possible pairs of SNPs in two genomic regions (genes) was the least effective. Second, the pairwise test almost had no power to detect interaction between two genomic regions (genes). Third, the effectiveness of the FRG-based test was substantially better than that of the pairwise tests. Fourth, the difference in power between the FRG and regression on PCA increases when the complexity of the interaction models increases.

To investigate the impact of sample size on the power, we plotted Figure 1E and Supplemental Figures 2 through 4, showing the power of three statistics for testing the interaction between two genomic regions (or genes) with only rare variants as a function of sample sizes under four interaction models, assuming 20% of the risk rare variants and the risk parameter $r = 1$ for Dominant OR Dominant and Recessive OR Recessive, and $r = 0.1$ for the Dominant AND Dominant and Threshold models, respectively. We observed similar power patterns of the three statistics under the four interaction models as those previously discussed. When sample sizes reach 10,000, the FRG model can be highly effective, but the effectiveness of the pairwise interaction test was still low even if the sample sizes increased to 10,000.

The FRG can also be applied to the presence of both common and rare variants. Figure 1F plotted the power curves of three sta-

Figure 1. (*A*) Power curves of three statistics: the FRG, the regression on PCA, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing the interaction between two genomic regions that consist of rare variants, for a quantitative trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$, under the Dominant OR Dominant model, assuming sample sizes of 2000. (*B*) Power curves of three statistics: the FRG, the regression on PCA, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing the interaction between two genomic regions that consist of rare variants, for a quantitative trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$, under the Dominant AND Dominant model, assuming sample sizes of 2000. (*C*) Power curves of three statistics: the FRG, the regression on PCA, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing the interaction between two genomic regions that consist of rare variants, for a quantitative trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$, under the Recessive OR Recessive model, assuming sample sizes of 2000. (*D*) Power curves of three statistics: the FRG, the regression on PCA, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing the interaction between two genomic regions that consist of rare variants, for a quantitative trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$, under the Threshold model, assuming sample sizes of 2000. (*E*) Power curves of three statistics: the FRG, the regression on PCA, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing the interaction between two genomic regions that consist of rare variants, for a quantitative trait as a function of the sample size at the significance level $\alpha = 0.05$, under the Dominant OR Dominant model, assuming the relative risk parameter $r = 0.1$. (*F*) Power curves of three statistics: the FRG, the regression on PCA, and pairwise interaction tests. Permutations were used to adjust for multiple testing, that is, for testing the interaction between two genomic regions with both common and rare variants, where 10% of the common variants and 10% of the rare variants were chosen as causal variants, as a function of the relative risk parameter r at the significance level $\alpha = 0.05$, under the Dominant OR Dominant model, assuming sample sizes of 2000.

tistics for testing interactions between two genomic regions (or genes) with both common and rare variants where 10% of the common variants and 10% of the rare variants were chosen as causal variants under the Dominant OR Dominant interaction model. Again, the FRG was the most effective among the three statistics. The power patterns of the tests for the interactions under the other three interaction models were similar. To limit the length of this publication, the investigation of the power of the tests in other scenarios is presented in Supplemental Note 5.

Application to real data examples

To further evaluate its performance, the FRG for testing interaction was applied to data from the NHLBI's ESP Project. The trait we considered was HDL. A total of 2225 individuals of European origin from 15 different cohorts in the ESP Project with no missing HDL phenotype value were included in the analysis. No evidence of cohort- and/or phenotype-specific effects or of other systematic biases was found (Tennessen et al. 2012). Exomes from related individuals were excluded from further analysis. The logarithm of HDL was taken as a trait value. The total number of genes tested for interactions, which included both common and rare variants, was 18,498. The remaining annotated human genes that did not contain any SNPs in our data set were excluded from the analysis. A P-value for declaring significant interactions after applying the Bonferroni correction for multiple tests was 4.75×10^{-10}. To examine the behavior of the FRG, we plotted QQ plots of the test

(Fig. 2). The QQ plots showed that the false-positive rate of the FRG for detection of interaction in some degree is controlled.

In total, 27 pairs of genes showed significant evidence of interaction with P-values $< 4.58 \times 10^{-10}$, which were calculated using the FRG model and logarithm transformation of the HDL. The results are summarized in Table 4 (below), where P-values for testing interactions between genes by regression on PCA and the minimum of P-values for testing all possible pairs of SNPs between two genes using a standard regression model are also listed. Since some complex traits in genetic studies often have non-normal distribution, we also used the rank-based inverse normal transformation (INT) to transform HDL (Beasley et al. 2009). The P-values for using INT of HDL with $c = \frac{1}{2}$ are included in Table 2. These 27 pairs of genes were derived from 35 genes. An additional 130 pairs of genes with P-values $< 9.87 \times 10^{-9}$ are listed in Supplemental Table 5.

Several remarkable features from these results were observed. First, we frequently observed the pairwise interaction between rare and rare variants (65.56%), and rare and common variants (34.44%). Less observed was the significant pairwise interaction between common and common variants with P-values for testing interactions $< 1.0 \times 10^{-6}$ in Table 2 and Supplemental Table 5, where variants with MAF < 0.05 are defined as rare variants, and variants with MAF ≥ 0.05 are defined as common variants. Second, pairs of SNPs between two genes jointly have significant interaction effects, but individually, each pair of SNPs makes mild contributions to the interaction effects, as shown in Table 3 and Supplemental Table 6. There were a total of 684 pairs of SNPs between genes *KCNK5* and *PRDM13*. Table 3 lists 35 pairs of SNPs with P-values < 0.0497. None of the 35 pairs of SNPs showed strong evidence of interaction. However, a number of pairs of SNPs between the genes *KCNK5* and *PRDM13* collectively demonstrated significant interaction. We observed similar interaction patterns in Supplemental Table 6, where eight pairs of SNPs between the genes *BHMT2* and *BMF* with P-values < 0.045 are listed. Third, the FRG often had a much smaller P-value to detect interaction than regression on the PCA and the minimum of P-values of pairwise tests had. Fourth, some investigators suggest that in genome-wide interaction analysis, only genes with large or mildly marginal genetic effects should be tested for interaction. However, we observed that genes may not show even mild marginal association, but they did demonstrate significant evidence of interaction (data were not shown). Fifth, computational time for gene-based interaction analysis is much less than that for pairwise tests. In Table 2, we tested a total of 27 pairs of genes and 9696 pairs of SNPs within them. The computational times for the FRG method for testing 27 pairs of gene interactions and by pairwise test for testing 9696 pairs of SNP interactions were 2.18 sec and 91.91 sec, respectively. The
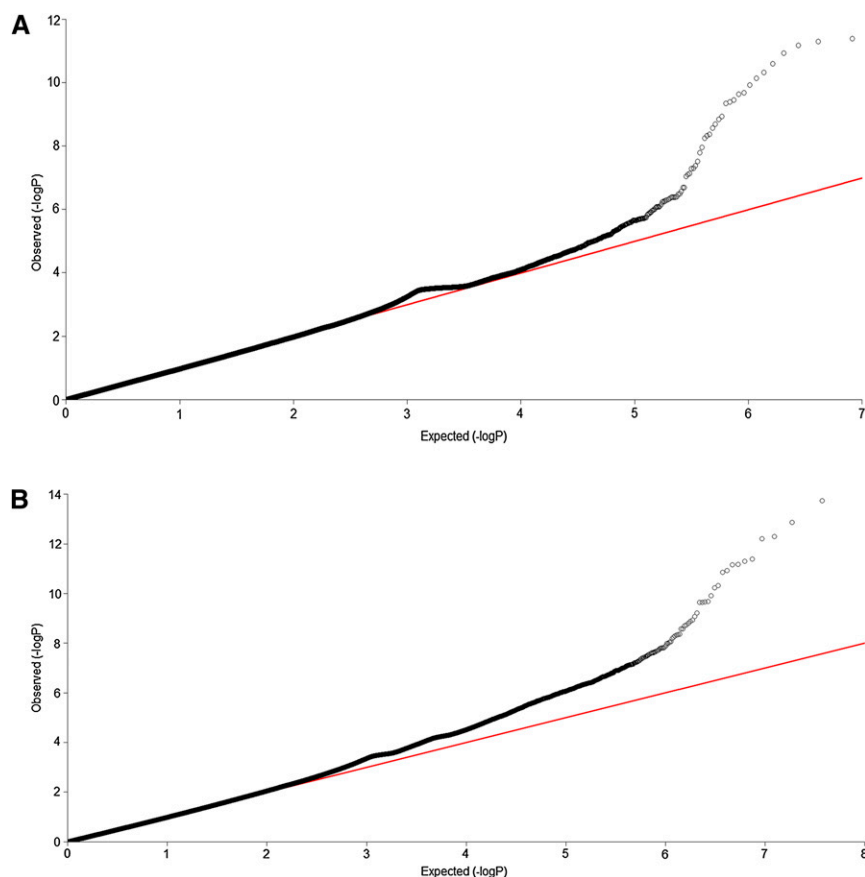

Figure 2. (*A*) QQ plot for the ESP data set. (*B*) QQ plot for the CHARGE-S data set.

Table 2. *P*-values of 27 pairs of significantly interacted genes identified by FRG

					P-values			
					FRG			
Chr	Gene1	Chr	Gene2	No. of Pairs	Original	INT	PCA	Pairwise (min)
5	GFM2	17	TBC1D3B	86	4.1×10^{-12}	6.1×10^{-11}	2.2×10^{-1}	5.1×10^{-13}
9	FAM206A	17	TBC1D3B	42	5.0×10^{-12}	1.1×10^{-11}	2.6×10^{-1}	1.2×10^{-12}
17	TBC1D3B	19	ADAT3	26	6.7×10^{-12}	1.8×10^{-10}	2.2×10^{-11}	8.2×10^{-12}
1	VSIG8	6	SLC35A1	493	7.0×10^{-12}	8.8×10^{-11}	9.5×10^{-2}	8.9×10^{-07}
6	VIP	17	TBC1D3B	24	1.2×10^{-11}	2.6×10^{-10}	1.2×10^{-12}	6.3×10^{-13}
1	CNKSR1	17	TBC1D3B	160	1.4×10^{-11}	8.6×10^{-11}	8.4×10^{-3}	1.6×10^{-11}
6	KCNK5	6	PRDM13	684	2.6×10^{-11}	2.5×10^{-09}	3.5×10^{-3}	5.1×10^{-07}
8	SNTB1	20	DBNDD2	1656	4.6×10^{-11}	1.8×10^{-06}	5.6×10^{-06}	2.7×10^{-05}
1	VSIG8	17	TBC1D3B	58	4.8×10^{-11}	1.7×10^{-10}	7.5×10^{-3}	1.1×10^{-06}
1	LRRC71	15	ITPKA	435	5.8×10^{-11}	7.8×10^{-10}	3.4×10^{-2}	1.1×10^{-09}
6	IRAK1BP1	6	PRDM13	324	7.1×10^{-11}	1.1×10^{-11}	7.1×10^{-06}	3.4×10^{-07}
5	TIGD6	9	C9orf91	500	1.2×10^{-10}	5.2×10^{-05}	7.1×10^{-2}	3.5×10^{-07}
6	GTPBP2	17	TBC1D3B	44	1.3×10^{-10}	3.6×10^{-10}	6.4×10^{-1}	9.4×10^{-13}
4	SCLT1	17	TBC1D3B	118	1.4×10^{-10}	2.5×10^{-09}	1.2×10^{-1}	2.9×10^{-12}
9	CTNNAL1	17	TBC1D3B	90	2.1×10^{-10}	1.8×10^{-10}	6.1×10^{-1}	1.6×10^{-06}
3	DIRC2	14	INSM2	320	2.2×10^{-10}	6.5×10^{-10}	9.4×10^{-04}	8.8×10^{-05}
8	SNTB1	18	DLGAP1-AS1	216	2.2×10^{-10}	3.6×10^{-05}	1.5×10^{-04}	2.7×10^{-05}
5	BHMT2	15	BMF	182	2.3×10^{-10}	2.0×10^{-05}	1.4×10^{-1}	5.3×10^{-08}
10	CYP2C18	17	TBC1D3B	72	2.3×10^{-10}	2.6×10^{-09}	1.1×10^{-10}	9.6×10^{-12}
6	PRDM13	14	ATP6V1D	288	2.7×10^{-10}	2.3×10^{-09}	4.1×10^{-09}	1.9×10^{-11}
14	HSPA2	18	C18orf56	115	2.9×10^{-10}	3.4×10^{-09}	4.3×10^{-1}	9.8×10^{-05}
6	PRDM13	17	TBC1D3B	36	3.5×10^{-10}	8.1×10^{-10}	1.6×10^{-09}	7.4×10^{-07}
3	GRIP2	6	PRDM13	2142	3.9×10^{-10}	6.4×10^{-10}	4.9×10^{-04}	2.2×10^{-07}
6	DEF6	6	MRPS18A	770	4.1×10^{-10}	5.4×10^{-05}	4.6×10^{-07}	3.1×10^{-06}
1	LRRC40	17	TBC1D3B	88	4.2×10^{-10}	2.2×10^{-09}	2.8×10^{-09}	1.7×10^{-07}
1	WASF2	6	PRDM13	540	4.5×10^{-10}	2.1×10^{-09}	2.2×10^{-08}	1.5×10^{-06}
2	CCDC115	14	CALM1	187	4.6×10^{-10}	2.8×10^{-06}	2.0×10^{-08}	2.6×10^{-07}

(INT) Rank-based inverse normal transformation.

computer configuration is as follows: CPU, Intel Core i7-3770 CPU at 3.4 GHz; memory (RAM), 16 GB. The interaction analysis by FRG on the entire set of genes was carried out on the cluster with 10 nodes, with each node having 24 cores (Intel Xeon CPU X5690 at 3.47 GHz). The running time for FRG on the entire set of genes was 18.3 h. Sixth, although interacting genes did not form large connected networks, we did observe some small interacted networks (Fig. 3). We observed three hub genes: *TBC1D3B*, *SNTB1*, and *PRDM13*. *TBC1D3B* had significant interactions with 12 genes (*P*-values < 4.20×10^{-10}) and interactions with 14 genes (*P*-values range from 9.10×10^{-9} to 5.10×10^{-10}). *SNTB1* strongly interacted with two genes and had modest interactions with another 26 genes (*P*-values varying from 9.19×10^{-9} to 9.10×10^{-10}). *PRDM13* strongly interacted with five genes (*P*-values < 4.58×10^{-10}) and had modest interactions with another 10 genes (*P*-values varying from 9.83×10^{-9} to 1.55×10^{-9}) (Table 2; Supplemental Table 5). *SNTB1* is a peripheral membrane protein. It is reported that *SNTB1* plays an essential role in regulating vascular tone and blood pressure (Lyssand et al. 2008). The multiple copies of *TBC1D3B* are located within a cluster of chemokine genes and might be a hominoid oncoprotein (Hodzic et al. 2006). We also observed modest interactions between *SNTB1* and *LDLR* (*P*-value < 4.76×10^{-7}) and between *SNTB1* and *LIPC* (*P*-value < 7.85×10^{-6}). *LDLR* and *LIPC* were reported to influence lipid levels in genome-wide association studies (GWAS) (Aulchenko et al. 2009). *PRDM13* is involved in transcriptional regulation (Chang et al. 2013). Point mutation in its strongly interacted gene *KCNK5* causes early-onset of autosomal dominant hypertension (Charmandari et al. 2012). Two interacted

subnetworks with *PRDM13* and *TBC1D3B* as hub genes were connected via directed interaction between two hub genes or via their interactions with *WASF2* and *EFNA3* (Fig. 3). *EFNA3* is a key regulator of embryogenesis and is expressed in human atherosclerotic plaque (Sakamoto et al. 2011). It was reported that *EFNA3* was a potential target of microRNA 210 as a novel therapy for treatment of ischemic heart disease (Hu et al. 2010).

To further evaluate the performance of the FRG for interaction analysis, we investigated whether 27 pairs of interacted genes (Table 2) in the ESP can be replicated in the CHARGE-S studies, which generated low-coverage, whole-genome sequencing data of 955 individuals from the ARIC (Atherosclerosis Risk in Communities), Framingham, and CHS (Cardiovascular Health Study) longitudinal cohorts after quality control with rich phenotypes including HDL cholesterol levels. A total of 25 pairs of genes in Table 2 in CHARGE-S were sequenced (*SNTB1* was not sequenced in CHARGE-S). Since we carried out 25 tests, the *P*-value for declaring replication after the Bonferroni correction for multiple tests was 0.002. We observed that 11 of the 25 pairs of significantly interacted genes (involving 14 genes) in the ESP project were replicated in the CHARGE-S study (Table 4). To further evaluate the performance of the FRG, we also considered a scenario where INT transformation of the HDL was taken as a trait value. The *P*-values for testing interactions between 10 pairs of genes selected from Table 2 using INT transformation of the HDL as a trait are included in Supplemental Note 6 (Supplemental Table S10). It is interesting to note that a subnetwork including six interactions with hub gene *PRDM13* and four interactions with hub gene *TBC1D3B* (Fig. 4) were replicated in the CHARGE-S study.

Table 3. *P*-values of 35 pairs of SNPs between genes *KCNK5* and *PRDM13* for testing interaction

GENE 1 *KCNK5* SNP1 (#RS)	GENE 2 *PRBM13* SNP2 (#RS)	*P*-value 2.55×10^{-11}
39159407	100056787	5.07×10^{-7}
39158947	100056787	2.39×10^{-5}
39196662	100056787	5.93×10^{-4}
39158843	100061914	6.10×10^{-3}
39159001	100061914	8.22×10^{-3}
39158816	100061914	9.78×10^{-3}
39196674	100061914	1.22×10^{-2}
39159072	100061914	1.48×10^{-2}
39196716	100061914	1.89×10^{-2}
39159346	100061914	1.98×10^{-2}
39196662	100061914	1.99×10^{-2}
39158836	100061914	2.15×10^{-2}
39158843	100060875	2.24×10^{-2}
39158843	100061120	2.71×10^{-2}
39159001	100060875	2.79×10^{-2}
39159407	100061914	2.81×10^{-2}
39159107	100061914	2.97×10^{-2}
39158843	100061150	3.32×10^{-2}
39163667	100061914	3.51×10^{-2}
39159001	100061120	3.63×10^{-2}
39158843	100054980	3.80×10^{-2}
39159379	100061120	3.89×10^{-2}
39159213	100061914	3.91×10^{-2}
39196662	100061221	3.94×10^{-2}
39158947	100061914	3.96×10^{-2}
39159379	100061914	4.03×10^{-2}
39158816	100061120	4.21×10^{-2}
39159001	100061150	4.34×10^{-2}
39163722	100061914	4.35×10^{-2}
39158843	100061153	4.39×10^{-2}
39159072	100061120	4.48×10^{-2}
39158843	100062380	4.63×10^{-2}
39159001	100054980	4.63×10^{-2}
39196674	100060875	4.97×10^{-2}

This again showed that *PRDM13* and *TBC1D3B* may make a large contribution to HDL-level variation.

Discussion

The current paradigm of pairwise interaction analysis was originally designed for testing the interaction for common variants and cannot be applied to genome-wide interaction analysis with rare variants due to its poor ability to detect interaction between rare variants, and rare and common variants, its prohibitive computational time, and the extremely large number of tests being conducted. To address these central themes and critical barriers in interaction analysis, we shift the paradigm of interaction analysis from the pairwise test to the collective group test, where we take a genome region (or gene) as a basic unit of interaction analysis and collectively test the interaction between all possible pairs of SNPs within two genome regions (or genes) and use FRG to develop a novel statistical framework for testing the interaction between two genomic regions (or genes). Using large simulations and real data analysis, we demonstrate the merits and limitations of the proposed new paradigm of interaction analysis.

The new approach uses all genetic information in the genome region to collectively test interaction between multiple SNPs within the regions. In the FRG approach to interaction analysis, we first expand the genotype function in a genomic region (gene) in terms of orthonormal basis functions. Genetic information across all variants in the genomic region, including all single variant variation and their linkage disequilibrium, is compressed into expansion coefficients. We use the compressed genetic information to globally test interaction between two genomic regions (genes). Therefore, the FRG for interaction analysis overcomes limitations inherent in pairwise interaction tests. By large simulations and real data analysis, we showed that the proposed FRG substantially increased the power and dramatically reduced the computational burden. In real data analysis, we also clearly demonstrate that pairs of SNPs between two genes jointly have significant interaction effects, but individually each pair of SNPs makes a mild contribution to interaction effects.

The pairwise interaction analysis is designed to test interactions between common and common variants, and is difficult to use to test interactions between rare and rare variants, and rare and common variants. There is an increasing need to develop statistics that can be used to test interactions among the entire allelic spectrum of variants. The FRG can efficiently test the interaction between rare and rare, rare and common, and common and common variants.

The essential problems in performing genome-wide interaction analysis in practice are the power of the test statistics, feasibility of computations, and efficient methods for *P*-value correction of multiple tests. Due to the lack of power of the widely used pairwise tests for interaction and its computational intensity,

Table 4. *P*-values of 11 pairs of genes that were significantly interacted in the ESP and CHARGE-S studies

|		Gene 1				Gene 2						*P*-values			
		No. of SNPs				No. of SNPs		No. of pairs		FRG		Pairwise (minimum)			
Chr	Symbol	ESP	CHARGE	Chr	Symbol	ESP	CHARGE	ESP	CHARGE	ESP	CHARGE	ESP	CHARGE		
6	VIP	12	74	17	TBC1D3B	2	19	24	1406	1.20×10^{-11}	5.00×10^{-04}	6.30×10^{-13}	2.20×10^{-04}		
6	KCNK5	38	347	6	PRDM13	18	99	684	34,353	2.60×10^{-11}	3.40×10^{-10}	5.10×10^{-07}	1.70×10^{-04}		
6	IRAK1BP1	18	248	6	PRDM13	18	99	324	24,552	7.10×10^{-11}	1.80×10^{-14}	3.40×10^{-07}	1.40×10^{-06}		
6	GTPBP2	22	52	17	TBC1D3B	2	19	44	988	1.30×10^{-10}	1.80×10^{-3}	9.40×10^{-13}	6.30×10^{-3}		
3	DIRC2	20	609	14	INSM2	16	47	320	28,623	2.20×10^{-10}	1.90×10^{-18}	8.80×10^{-05}	1.70×10^{-05}		
6	PRDM13	18	99	14	ATP6V1D	16	198	288	19,602	2.70×10^{-10}	1.20×10^{-3}	1.90×10^{-11}	7.10×10^{-10}		
6	PRDM13	18	99	17	TBC1D3B	2	19	36	1881	3.50×10^{-10}	4.20×10^{-04}	7.40×10^{-07}	1.80×10^{-3}		
3	GRIP2	119	445	6	PRDM13	18	99	2142	44,055	3.90×10^{-10}	5.00×10^{-12}	2.20×10^{-07}	1.40×10^{-06}		
6	DEF6	35	176	6	MRPS18A	22	116	770	20,416	4.10×10^{-10}	1.20×10^{-04}	3.10×10^{-06}	2.30×10^{-04}		
1	LRRC40	44	416	17	TBC1D3B	2	19	88	7904	4.20×10^{-10}	1.20×10^{-08}	1.70×10^{-07}	8.30×10^{-06}		
1	WASF2	30	564	6	PRDM13	18	99	540	55,836	4.50×10^{-10}	1.20×10^{-17}	1.50×10^{-06}	1.40×10^{-06}		
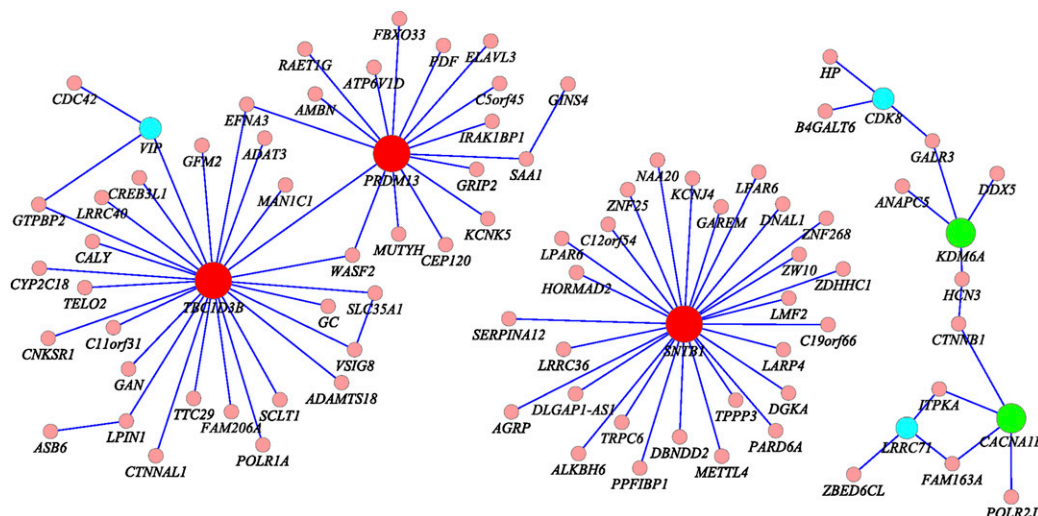
Figure 3. Networks of 27 pairs of genes showing significant evidence of interactions and genes showing mild interactions in Supplemental Table S5.

exploration of genome-wide gene–gene interactions has been limited (Ay 2002; Costanzo et al. 2010). Many geneticists question the universal presence of significant gene–gene interactions. Very few genome-wide interaction analyses with NGS data and very few results of significant interactions between rare and rare variants, and rare and common variants, have been reported. To our knowledge, we are among the first to conduct genome-wide interaction analysis with exome sequencing data. From genome-wide interaction analysis of HDL using the NHLBI's Exome Sequencing data, we have several important observations.

We observed that the majority of the significantly interacting genes showed no marginal association. Surprisingly, in 157 top pairs of interacting genes, the P-values for testing the marginal association of genes by the functional linear model ranged from 0.9933 to 0.00017. This strongly suggested that testing interactions for only genes with strong or mild marginal association will miss the majority of the interactions.

An interesting question to ask is what types of variants–rare or common–are more often present in the interaction. Our limited results showed that large proportions of interactions were due to interaction between rare and rare variants, and rare and common variants, but less significant pairwise interaction arose from interaction between common and common variants.

Whether interactions are most often present in isolation or whether interacting genes form networks is an open question. Our results indicated that interacting genes formed small interacting networks and that hub genes were present in the networks. These hub genes might be essential for interaction, which in turn may lead to important biological functions causing phenotype variation. We identified large networks that were generated from examining interactions between loci associated with serum lipid levels in recent GWAS, although interactions between genes in the networks were mild. We suspect that these genes in the network may jointly make contributions to the phenotype variation. Our preliminary results also showed that interactions can be replicated in two independent studies and observed that interactions with hub genes were more easily replicated.

It is well known that population stratification or cryptic relatedness may create artifactual LD, which in turn will lead to

spurious interaction. In the presence of population structure and cryptic relatedness, in general, we need to use mixed FRG to avoid the impact of population structure and cryptic relatedness on the tests. A detailed investigation is beyond the scope of this article.

NGS techniques generalize extremely high-dimensional genomic data. Transition of analysis from low-dimensional data to extremely high-dimensional data demands changes in the concept of interaction and quantitative trait models. Functional data analysis and the concept of group tests will provide a powerful tool for interaction analysis. However, the results presented in this article are considered preliminary. The number of basis functions in the expansion of genotype function will influence the performance of the FRG for interaction analysis. In practice, we select the number of
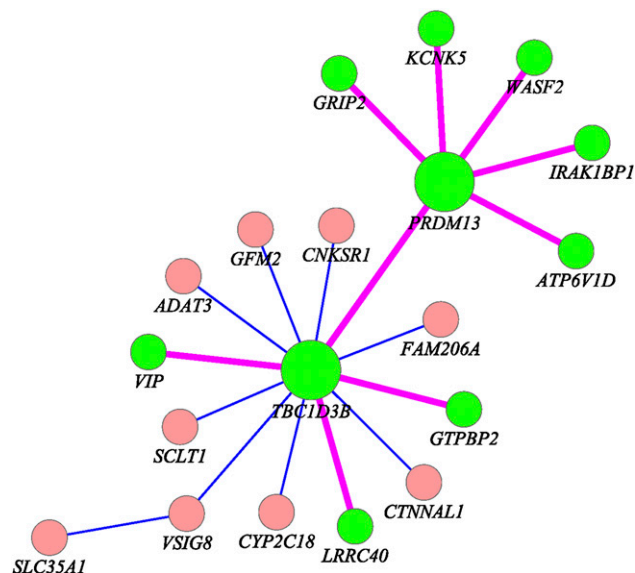

Figure 4. Nine interactions (pink color) between genes (green color) which form a subnetwork were replicated in the NHLBI's ESP and CHARGE-S studies.

basis functions which can explain 80%–90% of the genetic variation. Gene–gene interaction is an important but complex concept. Although functional data analysis and taking genomic regions as the unit of analysis can largely reduce the dimension of data for interaction analysis, genome-wide gene–gene interaction analysis still needs intensive computations. Genome-wide interaction analysis still poses great challenges. The main purpose of this article is to stimulate discussion about the optimal strategies for genome-wide interaction analysis. We hope that our results will greatly increase confidence in applying them to genome-wide gene–gene interaction analysis.

Data access

Software for implementing the proposed methods can be downloaded from Bioconductor (http://www.bioconductor.org/packages/2.14/bioc/html/FRGEpistasis.html) and our website http://www.sph.uth.tmc.edu/hgc/faculty/xiong/index.htm.

Acknowledgments

References

Ash RB, Gardner MF. 1975. *Topics in stochastic processes*. Academic Press, New York.

Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM. 2009. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41:** 47–55.

Ay N. 2002. Locality of global stochastic interaction in directed acyclic networks. *Neural Comput* **14:** 2959–2980.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11:** 773–785.

Beasley TM, Erickson S, Allison DB. 2009. Rank-based inverse normal transformations are increasingly used, but are they merited?. *Behav Genet* **39:** 580–595.

Bocianowski J. 2012. The use of weighted multiple linear regression to estimate QTL-by-QTL epistatic effects. *Genet Mol Biol* **35:** 802–809.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. *Nature* **490:** 535–538.

Chang JC, Meredith DM, Mayer PR, Borromeo MD, Lai HC, Ou Y, Johnson JE. 2013. Prdm13 mediates the balance of inhibitory and excitatory neurons in somatosensory circuits. *Dev Cell* **25:** 182–195.

Charmandari E, Sertedaki A, Kino T, Merakou C, Hoffman DA, Hatch MM, Hurt DE, Lin L, Xekouki P, Stratakis CA, et al. 2012. A novel point mutation in the KCNJ5 gene causing primary hyperaldosteronism and early-onset autosomal dominant hypertension. *J Clin Endocrinol Metab* **97:** E1532–E1539.

Chen Z, Cui W. 2010. A two-phase procedure for QTL mapping with regression models. *Theor Appl Genet* **121:** 363–372.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10:** 392–404.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. *Science* **327:** 425–431.

Ferraty F, Romain Y. 2010. *The Oxford handbook of functional data analysis*. Oxford University Press, New York.

Fisher R. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* **52:** 399–433.

Guyon I, Elisseeff A. 2003. An introduction to variable selection and feature selection. *J Mach Learn Res* **3:** 1157–1182.

Guyon I, Gunn S, Nikravesh M, Zadeh L. 2006. *Feature extraction: foundations and applications* (Studies in fuzziness and soft computing). Springer, New York.

Hodzic D, Kong C, Wainszelbaum MJ, Charron AJ, Su X, Stahl PD. 2006. TBC1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17q12. *Genomics* **88:** 731–736.

Hu S, Huang M, Li Z, Jia F, Ghosh Z, Lijkwan MA, Fasanaro P, Sun N, Wang X, Martelli F, et al. 2010. MicroRNA 210 as a novel therapy for treatment of ischemic heart disease. *Circulation* **122:** S124–S131.

Joyce P, Tavaré S. 1995. The distribution of rare alleles. *J Math Biol* **33:** 602–618.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83:** 311–321.

Li J, Tang R, Biernacka JM, de Andrade M. 2009. Identification of gene-gene interaction using principal components. *BMC Proc* **3:** S78.

Luo L, Boerwinkle E, Xiong MM. 2011. Association studies for next-generation sequencing. *Genome Res* **21:** 1099–1108.

Luo L, Zhu Y, Xiong MM. 2012. Quantitative trait locus (QTL) analysis for next-generation sequencing with the functional linear models. *J Med Genet* **49:** 513–524.

Lyssand JS, DeFino MC, Tang XB, Hertz AL, Feller DB, Wacker JL, Adams ME, Hague C. 2008. Blood pressure is regulated by an α_{1D}-adrenergic receptor/dystrophin signalosome. *J Biol Chem* **283:** 18792–18800.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5:** e1000384.

Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean St P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337:** 100–104.

Prabhu S, Pe'er I. 2012. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res* **22:** 2230–2240.

Ramsay JO, Silverman BW. 2005. *Functional data analysis*. Springer, New York.

Saeys Y, Inza I, Larranaga P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23:** 2507–2517.

Sakamoto A, Sugamoto Y, Tokunaga Y, Yoshimuta T, Hayashi K, Konno T, Kawashiri MA, Takeda Y, Yamagishi M. 2011. Expression profiling of the ephrin (EFN) and Eph receptor (EPH) family of genes in atherosclerosis-related human cells. *J Int Med Res* **39:** 522–527.

Steen KV. 2011. Travelling the world of gene–gene interactions. *Brief Bioinform* **13:** 1–19.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337:** 64–69.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89:** 82–93.