Xu et al. BMC Genomics (2017) 18:385
DOI 10.1186/s12864-017-3777-4

BMC Genomics

METHODOLOGY ARTICLE Open Access

Functional regression method for whole genome eQTL epistasis analysis with sequencing data

Kelin Xu[1,2], Li Jin[1] and Momiao Xiong[1,3,4*]

Abstract

Background: Epistasis plays an essential rule in understanding the regulation mechanisms and is an essential component of the genetic architecture of the gene expressions. However, interaction analysis of gene expressions remains fundamentally unexplored due to great computational challenges and data availability. Due to variation in splicing, transcription start sites, polyadenylation sites, post-transcriptional RNA editing across the entire gene, and transcription rates of the cells, RNA-seq measurements generate large expression variability and collectively create the observed position level read count curves. A single number for measuring gene expression which is widely used for microarray measured gene expression analysis is highly unlikely to sufficiently account for large expression variation across the gene. Simultaneously analyzing epistatic architecture using the RNA-seq and whole genome sequencing (WGS) data poses enormous challenges.

Methods: We develop a nonlinear functional regression model (FRGM) with functional responses where the position-level read counts within a gene are taken as a function of genomic position, and functional predictors where genotype profiles are viewed as a function of genomic position, for epistasis analysis with RNA-seq data. Instead of testing the interaction of all possible pair-wises SNPs, the FRGM takes a gene as a basic unit for epistasis analysis, which tests for the interaction of all possible pairs of genes and use all the information that can be accessed to collectively test interaction between all possible pairs of SNPs within two genome regions.

Results: By large-scale simulations, we demonstrate that the proposed FRGM for epistasis analysis can achieve the correct type 1 error and has higher power to detect the interactions between genes than the existing methods. The proposed methods are applied to the RNA-seq and WGS data from the 1000 Genome Project. The numbers of pairs of significantly interacting genes after Bonferroni correction identified using FRGM, RPKM and DESeq were 16,2361, 260 and 51, respectively, from the 350 European samples.

Conclusions: The proposed FRGM for epistasis analysis of RNA-seq can capture isoform and position-level information and will have a broad application. Both simulations and real data analysis highlight the potential for the FRGM to be a good choice of the epistatic analysis with sequencing data.

Keywords: Gene-gene interaction, Multivariate functional regression, Functional regression models, RNA-seq, Next-generation sequencing, Association studies, eQTL

* Correspondence: momiao.xiong@uth.tmc.edu
[1]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200438, China
[3]Department of Biostatistics, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
Full list of author information is available at the end of the article

Background

Epistatic effect in gene expression, defined as the departure from additive effects in a linear model of eQTL analysis [1], plays an essential role in understanding the gene regulation and disease mechanisms [2–4]. One polymorphism's effect on expression of a gene depends on other polymorphisms present in the genome [5]. Epistasis analysis of gene expressions will substantially improve the understanding of the genetic architecture of gene expression and facilitate mechanistic insights into complex traits [6, 7]. However, eQTL epistasis analysis remains fundamentally unexplored due to large computational challenges and data availability [6].

Gene expression is an intermediate phenotype that bridges the genotype and higher level phenotypes such as diseases [8, 9]. Studying the effect of epistasis on the gene expression could provide a better understanding of the genetic architecture and gene regulation. The importance of detecting the epistatic effect on the gene expression has been emphasized in many recent studies [10, 11]. However, the corresponding methods are relatively rare. The widely used statistical methods for identifying eQTL epistasis are designed for microarray expression data where an overall expression of the gene is taken as a quantitative trait and all methods for QTL epistasis analysis can be used for eQTL epistasis analysis [10, 12].

Application of next generation sequencing (NGS) techniques to the genetic analysis of gene expression involves (1) generating millions of short reads of mRNA or cDNA which are mapped to the genome and lead to a sequence of read counts at the hundreds of millions of genomic positions [13–16] and (2) generating millions or 10 millions of genetic variants. RNA-seq counts vary greatly across the gene [17]. Count variations can be due to experimental bias such as fragmentation methods, reverse-transcription [16], sequence-specific bias and sequencing technology variation [18]. However, count variation can also be caused by variation in splicing, transcription start sites, polyadenylation sites, post-transcriptional RNA editing across the entire gene, and transcription rates of the cells [13–16, 18]. RNA-seq data can be viewed as a function or a curve of the genomic position and hence can be taken as a function-valued trait.

Although RNA-seq data are measured as a function, the widely used methods for genetic studies of the RNA-seq in humans are the same as that for the traditional single-valued quantitative traits where a single number for overall expression of the gene is taken as a quantitative trait. These methods use summary statistics to measure or represent gene expressions assayed by NGS techniques and cannot capture the expression variations across the gene due to splicing, transcription start sites, polyadenylation sites, post-transcriptional RNA editing across the entire gene, and transcription rates of the various cells. The summary statistic-based epistasis analysis of the RNA-seq fails to utilize all transcripts information.

The critical barrier in epistasis analysis is to deal with rare variants. The traditional statistical methods for epistasis analysis were originally designed for testing the interaction between common variants and are difficult to apply to rare variants due to high type 1 error rates, severe multiple testing, prohibitive computational time and low power [19]. Whole genome RNA-req eQTL analysis poses a significant challenge. To meet the challenge, we developed a nonlinear functional regression model (FRGM) with functional responses where the position-level read counts within a gene are taken as a function of genomic position, and functional predictors where genotype profiles are viewed as a function of genomic position, for epistasis analysis with RNA-seq data, which allows simultaneous capture of all space information hidden in the RNA-seq data and genetic variation data, but with substantially reduced dimensions. Instead of testing the interaction of all possible pair-wises SNPs, the FRGM takes a gene as a basic unit for epistasis analysis, which tests for the interaction of all possible pairs of genes and uses all the information that can be accessed to collectively test interaction between all possible pairs of SNPs within two genome regions (or genes). The proposed FRGM for epistasis analysis of the RNA-seq can capture isoform and position-level information and will have a broad application.

The FRGM for epistasis analysis has several remarkable features. First, the FRGM accounts for the change in the position-level read counts, while preserving the intrinsic structure and all the positional-level genetic information. Second, the FRGM simultaneously utilizes both correlation information among the RNA-seq at different genomic positions and among all variants in a genomic region. Third, the multicollinearity problems in the FRGM which may be presented in both the RNA-seq and genetic variation are alleviated. Fourth, the FRGM expands both position-level read count function and genotype function in terms of orthogonal eigenfunction, which leads to substantial dimension reduction in both RNA-seq data and SNP data. The FRGM for epistasis analysis of function-valued traits which capture key information in the data is expected to open a new route for epistasis analysis of RNA-seq data.

To evaluate its performance for epistasis analysis of the RNA-seq, we use large scale simulations to calculate the type I error rates and evaluate the power of the proposed FRGM for detecting epistasis. To further evaluate its performance, the FRGM for epistasis analysis is applied to 350 samples with both RNA-seq and NGS data from the 1000 Genomes Project. An R packge for implementing the

developed FRGM for epistasis analysis of RNA-seq and NGS data can be downloaded from our website https://sph.uth.edu/research/centers/hgc/xiong/software.htm.

Results

Null distribution of test statistics

To examine the null distribution of test statistics, we performed a series of simulation studies to compare their empirical levels with the nominal ones. We consider three models for type 1 error rate simulations: model 1 with no marginal effects, model 2 with marginal effects at the first gene and model 3 with marginal effects at both the first and second genes.

We generated 100,000 chromosomes by resampling from the 350 European samples with genetic variants in five genes*: IRAK3, ACSS3, SUV420H1, ETV7*, and *HPS4* from the next generation sequencing data in the 1000 Genomes Project. The summary statistics of the variants in five genes are summarized in Additional file 1: Table S1. The marginal genetic effects will be estimated from the data. 100 genes with RNA-seq data were randomly selected from GEUVADIS project. They were used to develop the models for generating RNA-seq data in simulation (Detailed description were referred to Method Section).10 pairs of genes were selected from five genes : IRAK3, ACSS3, SUV420H1, ETV7, and HPS4 with genotype data from 1000 Genome Project dataset.

The number of sampled individuals from the population ranged from 1000 to 5,000, and 5,000 simulations were repeated. We randomly selected 10% of the SNPs as causal variants from five genes: *IRAK3, ACSS3, SUV420H1, ETV7*, and *HPS4*. We perfume gene-gene interaction tests for 10 pairs of genes selected from five genes with genotypes under the three models for 5000 times. The type 1 error rates were averaged over 10 pairs of genes with genotype data and 5,000 simulations for each model. Tables 1, 2 and 3 summarized the type I error rates of the test statistics for testing the interaction between two genes with no marginal effect, marginal effect at the first gene and marginal effects at both genes consisting only of rare variants and both common and rare variants, respectively, averaged over 100 genes with RNA-seq data and 10 pairs of genes with genotype data at the nominal levels $\alpha = 0.05$, $\alpha = 0.01$ and $\alpha = 0.001$. These results clearly showed that the type I error rates of the FRGM-based test statistics for testing interaction between two genes with or without marginal effects were not appreciably different from the nominal α levels.

Table 1 Average type 1 error rates of the statistic for testing interaction between two genes with no marginal effect over 10 pairs of genes

	Rare Variants			Common & Rare Variants		
Sample Size	0.05	0.01	0.001	0.05	0.01	0.001
1000	0.0495	0.0099	0.0009	0.0497	0.0101	0.0010
2000	0.0501	0.0094	0.0010	0.0510	0.0100	0.0011
3000	0.0475	0.0097	0.0011	0.0498	0.0103	0.0011
4000	0.0497	0.0097	0.0011	0.0501	0.0101	0.0009
5000	0.0499	0.0104	0.0011	0.0511	0.0108	0.0010

Table 2 Average type 1 error rates of the statistic for testing interaction between two genes with marginal effect at the first genes over 10 pairs of genes

Rare Variants			Common & Rare Variants		
0.05	0.01	0.001	0.05	0.01	0.001
0.0507	0.0094	0.0007	0.0491	0.0100	0.0010
0.0500	0.0098	0.0009	0.0489	0.0099	0.0012
0.0508	0.0108	0.0010	0.0496	0.0096	0.0011
0.0490	0.0101	0.0010	0.0498	0.0103	0.0011
0.0490	0.0101	0.0010	0.0506	0.0093	0.0008

Power evaluation

To evaluate the performance of the functional regression model for testing the epistatic effect on gene expression, we estimated the power through simulations. We generated 100,000 chromosomes by resampling from the 350 European samples with genetic variants in two genes: IRAK3 and ACSS3 from the next generation sequencing data in 1000 Genomes Project. We randomly selected 20% variants as causal variants, assumed that there were k_1 SNPs in the first gene, and k_2 SNPs in the second gene. Two thousand individuals were sampled. We assumed that both marginal effects and epistasis effects were a function of the genomic position and used the multiple regression models to generate the RNA-seq data under four interaction models: Dominant OR Dominant, Dominant AND Dominant, Recessive OR Recessive and Threshold (See the Methods section).

We compared the power of the FRGM with both functional response and functional predictors (BFGM), FRGM with scalar response and functional predictors (SFGM) and regression on principal component analysis (PCA). For the PCA method, the PCA was performed on the RNA-seq data and the number of PCs were

Table 3 Average type 1 error rates of the statistic for testing interaction between two genes with marginal effects at two genes over 10 pairs of genes

	Rare Variants			Common & Rare Variants		
Sample Size	0.05	0.01	0.001	0.05	0.01	0.001
1000	0.0501	0.0108	0.0011	0.0486	0.0103	0.0010
2000	0.0493	0.0098	0.0010	0.0495	0.0101	0.0010
3000	0.0499	0.0095	0.0011	0.0497	0.0101	0.0011
4000	0.0494	0.0099	0.0010	0.0496	0.0096	0.0008
5000	0.0489	0.0097	0.0011	0.0497	0.0107	0.0010

selected to explain 80% variance of number of reads at different genomic positions. The multiple functional regression was performed to analyze the data [20]. In the BFGM, both RNA-seq and genotype profiles were taken as a function of genomic position and expanded in terms of functional principal components.

Figure 1a-d plotted the power curves of three statistics: BFGM, SFGM and PCA to test the interaction between two genes with rare variants under the Dominant OR Dominant, Dominant AND Dominant, Recessive OR Recessive and Threshold models, respectively. In the simulation, 20% of the rare variants were randomly selected as the causal variants. These power curves were a function of the risk parameter at the significance level $\alpha = 0.05$. We observed that under all four interaction models the BFGM had the highest power, followed by the regression on PCA. Power of the SFGM was the lowest. The results demonstrated that summary statistics such as RPKM for measuring gene expression could not capture the expression variations across the gene and almost had no power to detect the interaction between two genes with rare variants.

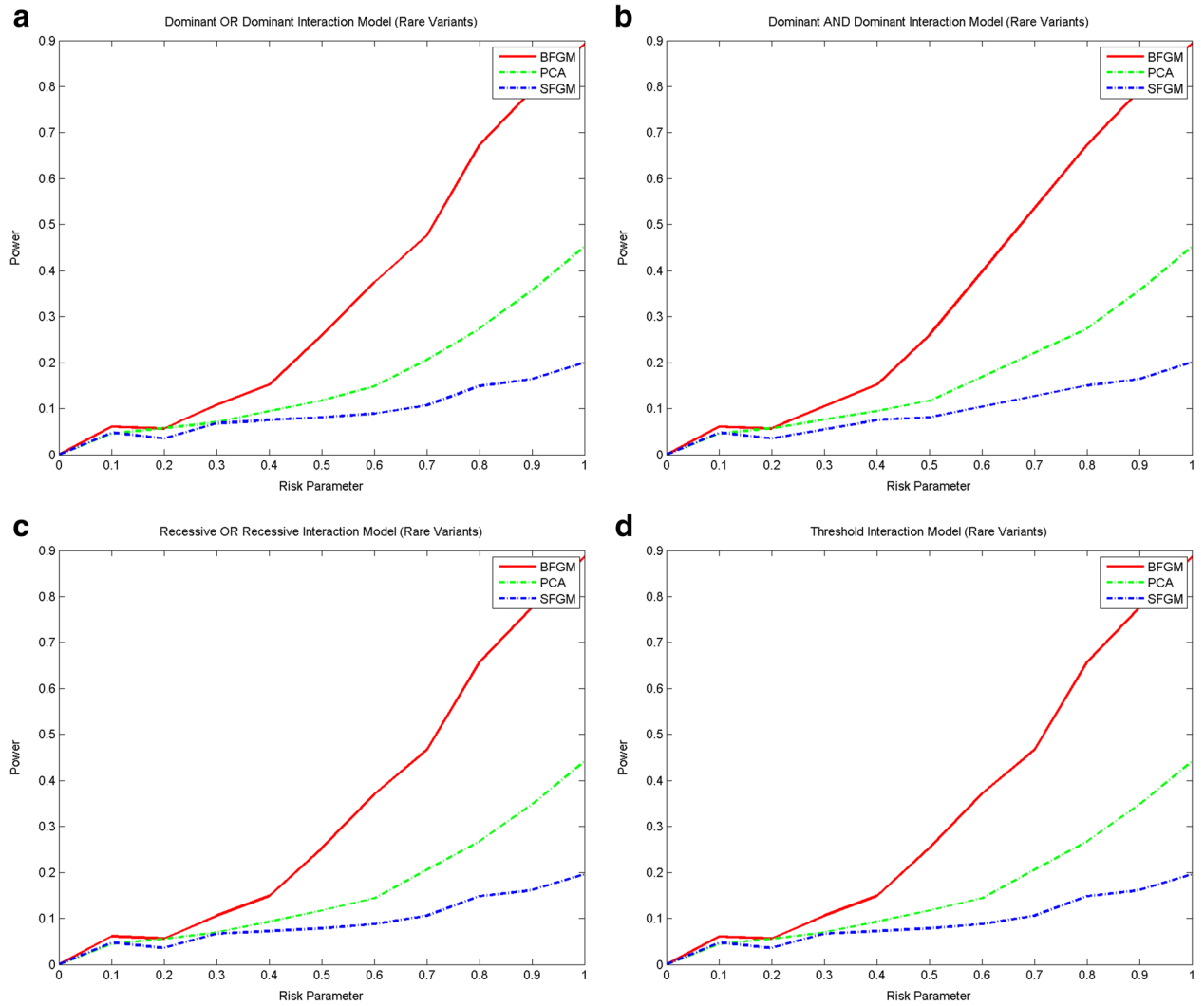

Fig. 1 a. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of rare variants with the RNA-seq trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$ under the Dominant OR Dominant model, assuming sample sizes of 2,000. **b**. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of rare variants with RNA-seq trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$ under the Dominant AND Dominantmodel, assuming sample sizes of 2,000. **c**. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of rare variants with RNA-seq 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 2,000. **d**. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of rare variants with RNA-seq 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 2,000

The BFGM can also be applied to the presence of both common and rare variants. Figure 2a-d plotted the power curves of three statistics for testing interaction between two 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, Dominant AND Dominant, Recessive OR Recessive and Threshold models, respectively. The power patterns of tests for the interactions between two genes with both common and rare variants were similar to that with rare variants only. The BFGM had the highest power, followed by the PCA and the SFGM. However, we noticed that the power of the SFGM for epistasis analysis in the presence of common variants increased substantially. Under some models such as the Dominant OR Dominant model, the SFGM would have enough power to detect interactions between two genes with common variants.

RNA-seq data and NGS data

The BFGM was applied to the RNA-seq data in the GEUVADIS RNA Sequencing Project [21] and the WGS data in the1000 Genomes Project. A total of 350 samples

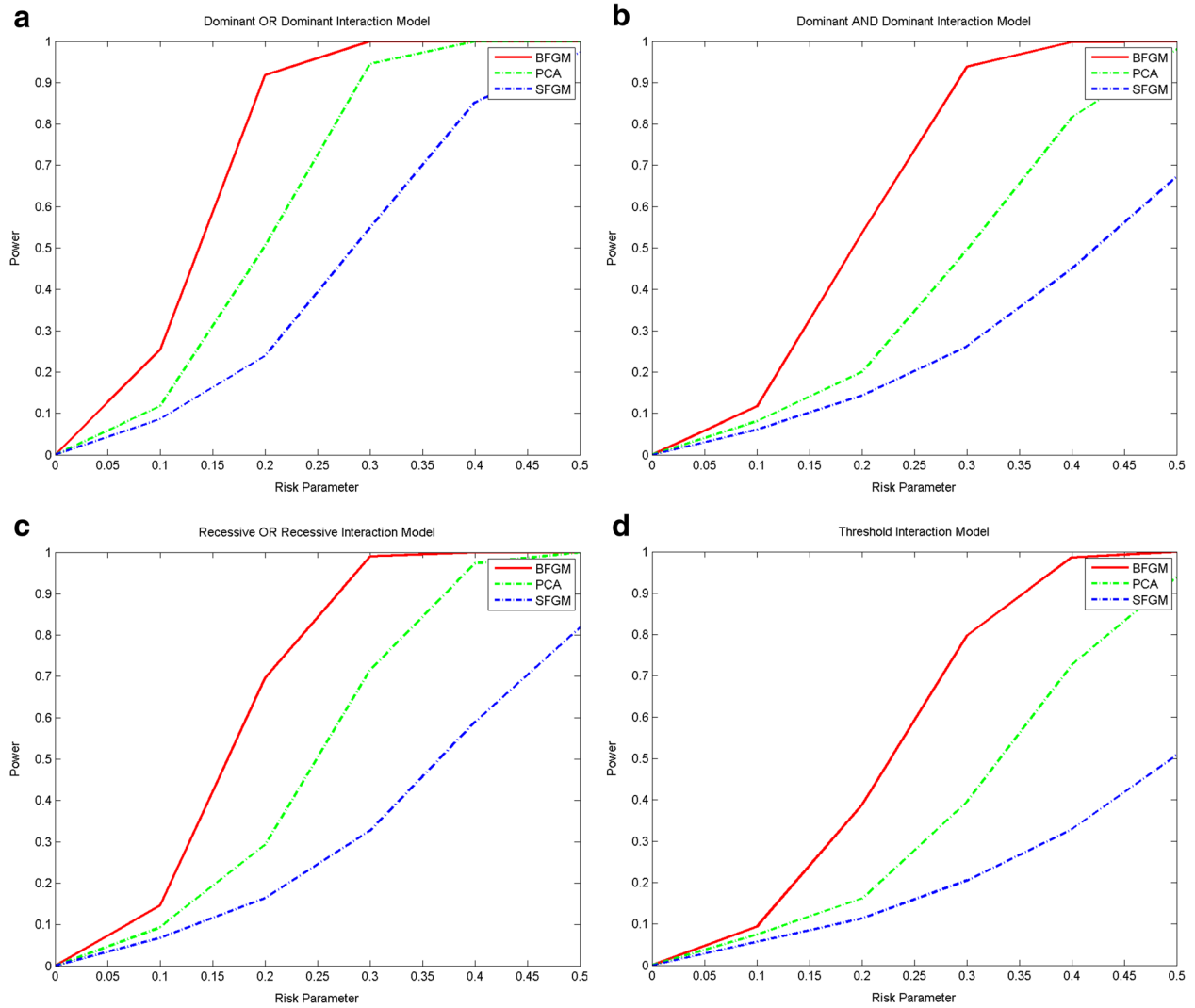

Fig. 2 a. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of both common and rare variants with the RNA-seq trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$ under the Dominant OR Dominant model, assuming sample sizes of 2,000. **b**. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of both common and rare variants with RNA-seq trait as a function of the relative risk parameter r at the significance level $\alpha = 0.05$ under the Dominant AND Dominantmodel, assuming sample sizes of 2,000. **c**. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of both common and rare variants with RNA-seq 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 2,000. **d**. Power curves of three statistics: the BFGM, regression on PCA, SFGM, for testing interaction between two genomic regions that consist of both common and rare variants with RNA-seq 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 2,000

with European origin was shared between the GEUVADIS RNA Sequencing Project and 1000 Genomes Project, which had combined transcriptome (22,706 gene expressions measured by RNA-seq) and genome sequencing data (2,708,453 SNPs in 24,519 genes). After removing singleton SNPs, repeated SNPs, and filtering out the SNPs violating HW equilibrium [22] (P value $< 10^{-9}$ for declaring HW disequilibrium), 2,566,261 SNPs in 18,986 genes were included in the epistasis analysis. In the RNA-seq data pre-processing, we removed the genes whose expressing rates were less than 30% and the genes that did not contain any SNPs. Finally, RNA-seq data of the 15,656 genes were included in the analysis. We used DESeq [23] to normalize the RNA-seq data.

Cis-trans interactions

We considered the RNA-seq curve of the target gene as a function-valued trait. The target gene selected from the 15656 gene expressions was referred to as gene 1. We selected one of the remaining 18985 genotyping genes as gene 2. We used BFGM to test for the interactions between gene 1 and gene 2 influencing the expression of the target gene 1. The total number of gene pairs tested for interactions which included both common and rare variants was 297,229,160. A P-value for declaring significant interaction after applying the Bonferroni correction for multiple tests was 1.68×10^{-10}. To examine the behavior of the BFGM, we plotted the QQ plot of the test (Fig. 3). QQ plot showed that the false positive rate of the BFGM for detection of epistasis was controlled.

For comparisons, the SFGM was also applied to the dataset. RPKM and DESeq were used to compute the overall expression value of genes from the RNA-seq data. All the expression values were processed by the rank-based inverse normal transformation [24]. For both common and rare variants, in total, 162361, 260 and 51 significant *cis*-trans interactions regulating the gene expressions were identified by the BFGM, SFGM with the RPKM and DESeq, respectively. We observed 9,846genes whose expressions were influenced by 16,2361*cis*-trans interactions. We found that the average number of epistasis influencing each gene was 16. A total of 3,505 gene expressions were influenced by one significant *cis*-trans gene-gene interactions, 169 gene expressions were influenced by more than 100 *cis*-trans gene-gene interactions. Figure 4 presented a histogram that showed a distribution of the *cis*-trans gene-gene interactions.

The P-values of the top 20 interactions between genes ranked by the BFGM method were summarized in Table 4 where P-values for testing interactions between genes by the SFGM (RPKM, DESeq and RNAmin) and min P-values were also listed. The RNAmin denoted the minimum of P-values computed by the SFGM method with the number of reads at each genome position of the gene as the scalar response in the functional regression model. The min P-values denoted to take the minimum of all P-values for testing all possible pairs of SNPs between two genes using functional regression model with functional response and scalar predictors. Table 4 showed several remarkable features. First, we often observed the pair-wise interaction between rare and rare variants (34.38%), and rare and common variants (59.38%). Less observed was the significant pair-wise interaction between common and common variants (6.25%). Second, significant interactions between two

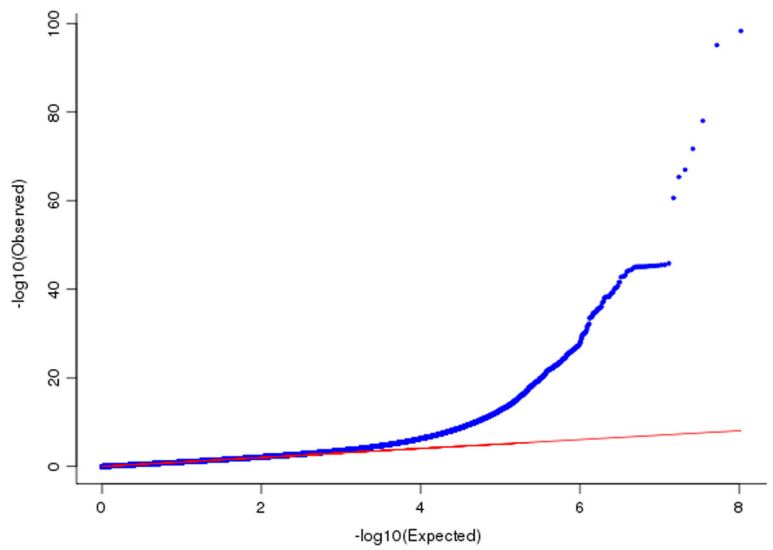

Fig. 3 QQ plot of P-values from the BFGM for testing the *cis*-trans interactions between two genes influencing transcription

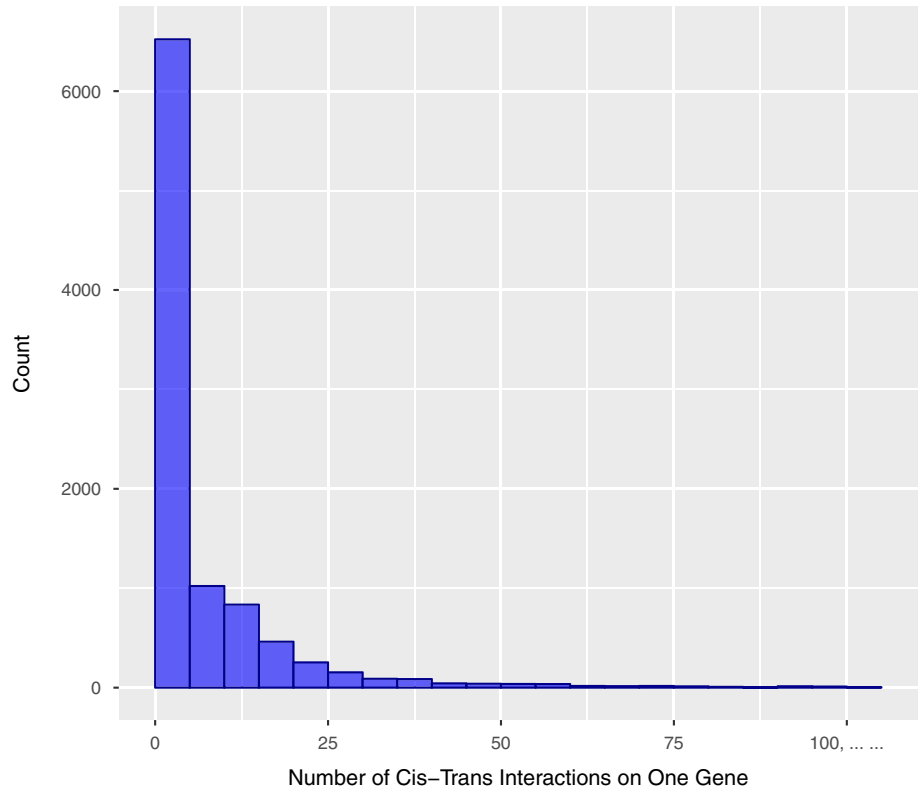

Fig. 4 A histogram showing a distribution of the number of cis-trans gene-gene interactions on each gene expression

Table 4 *P*-values of top 20 genes ranked by the BFGM methods

							P-value (Interaction)				
								SFGM			
Gene Expression	Gene 1	Chr	Marginal (eQTL)	Gene 2	Chr	Marginal (eQTL)	BFGM	RPKM	DESeq	RNA-min	min *P*-value
ULK4	ULK4	3	1.45E-01	C19orf70	19	7.96E-06	0.00E + 00	4.17E-01	4.98E-01	0.00E + 00	0.00E + 00
ULK4	ULK4	3	1.45E-01	OR10A2	11	6.04E-15	4.07E-305	4.33E-05	4.97E-03	0.00E + 00	0.00E + 00
CCDC13	CCDC13	3	5.60E-01	TMEM121	14	9.94E-24	2.91E-302	4.98E-03	1.73E-02	1.36E-304	2.12E-12
ULK4	ULK4	3	1.45E-01	PSMC5	17	7.30E-23	5.72E-267	2.79E-05	1.50E-03	0.00E + 00	0.00E + 00
ULK4	ULK4	3	1.45E-01	COX5B	2	1.15E-03	2.46E-242	1.82E-03	3.06E-02	2.46E-259	4.94E-323
NKX2-5	NKX2-5	5	3.41E-01	TP53TG3D	16	5.81E-10	2.18E-226	6.66E-02	7.61E-02	1.15E-228	0.00E + 00
ASIC2	ASIC2	17	2.24E-02	RPS16P5	6	4.08E-05	3.39E-226	1.12E-01	7.77E-02	1.76E-158	8.01E-237
TMEM132E	TMEM132E	17	8.25E-02	LOC100144602	4	9.50E-51	2.04E-213	1.34E-01	1.30E-01	2.27E-142	1.14E-215
TMEM98	TMEM98	17	6.66E-01	LOC100144602	4	9.28E-51	4.89E-213	8.10E-02	1.13E-01	4.06E-144	6.05E-216
SPACA3	SPACA3	17	7.13E-02	LOC100144602	4	1.41E-50	9.18E-211	9.72E-03	1.21E-02	8.78E-141	4.90E-214
ASIC2	ASIC2	17	2.24E-02	OR5B12	11	3.09E-05	2.89E-205	1.97E-01	1.18E-01	3.63E-141	1.18E-259
CCL1	CCL1	17	6.53E-02	TINF2	14	3.57E-22	3.85E-205	1.23E-01	1.31E-01	1.52E-157	3.33E-210
SCN2A	SCN2A	2	1.41E-01	DEFB4B	8	1.30E-32	5.53E-203	1.88E-02	4.91E-02	1.30E-104	2.67E-236
ZNF254	ZNF254	19	4.41E-01	OR2V1	5	2.20E-13	1.19E-183	2.35E-02	4.91E-02	5.50E-225	2.88E-259
KRT5	KRT5	12	2.45E-02	OR5K1	3	1.03E-28	3.26E-177	1.15E-02	2.33E-02	2.29E-201	5.79E-199
FNDC8	FNDC8	17	4.83E-01	LOC100144602	4	8.13E-49	3.46E-172	4.38E-03	6.19E-02	1.20E-124	8.03E-175
CCT6B	CCT6B	17	3.25E-01	LOC100144602	4	8.81E-49	4.24E-172	7.49E-03	3.59E-02	3.10E-125	1.72E-177
TMEM163	TMEM163	2	1.16E-01	HIST1H4H	6	1.79E-09	8.50E-170	6.74E-02	4.79E-02	3.40E-112	1.19E-25
ASIC2	ASIC2	17	2.24E-02	LOC100144602	4	9.66E-51	6.06E-165	9.22E-01	8.72E-01	3.82E-104	9.50E-236
KRT5	KRT5	12	2.45E-02	IFNA7	9	3.70E-16	2.43E-163	2.26E-02	1.83E-02	2.15E-178	2.14E-211

genes often indicated that at least one significant pair of SNPs in two genes could be observed (min *P*-values were small). However, we can observe that pairs of SNPs between two genes jointly had significant interaction effects, but individually each pair of SNPs mildly contributed to the interaction effects. Third, the BFGM often had a much smaller *P*-value to detect interaction than other tests. Fourth, we observed that genes may not show even mild marginal association, but they did demonstrate significant evidence of interaction. If only the interactions between two marginally significant genes are tested, some significant interactions may be missed. The fifth, the BFGM tremendously reduced computation burden.

To further assess the validity of the BFGM for epistasis analysis with RNA-seq data, we randomly selected six pairs of genes from the significant 162361 gene-gene interactions. The *P*-values for testing the interactions of six pairs of genes using the BFGM and SFGM were summarized in Table 5. Table 5 showed that six significant interactions identified by the BFGM significantly influenced read count variation at least at one genomic position within the gene. To explain why the BFGM had higher power to detect interaction than the SFGM, we presented Fig. 5a-f showing the RNA-seq profiles and overall expression level of the genes *PLA2G4A, PLA2G6, PLAUR, PLD4, PLD6* and *PLEKHA3* of two individuals, respectively. These figures showed that the overall expression levels of the individuals were the same, but their RNA-seq profiles were quite different. This demonstrated that unlike the RNA-seq profiles, the overall expression levels cannot capture the expression variation across the genes. Therefore, the SFGM using summary statistics as a trait will have less power to detect the interaction than the BFGM using the RNA-seq profiles as a function-valued trait.

To investigate whether the top 20 interactions were caused by the linkage disequilibrium (LD) or not, we listed the maximum of r^2 between all possible SNPs in the top 20 significantly interacting pairs of genes and the *P*-values for testing their presence of LD in Table 6. We did not observe the strong LD between the interacting genes.

Table 5 The *P*-values of randomly selected 6 pairs of genes from the significant 162361 gene-gene interactions

GENE1	GENE2	*P*-values			
		BFGM	SFGM		
			RPKM	DESeq	RNAmin
PLA2G4A	OR7E2P	1.18E-18	3.68E-02	2.79E-02	1.68E-28
PLA2G6	FGF14-AS2	2.43E-14	8.19E-01	4.90E-01	1.42E-26
PLAUR	CAPNS2	5.14E-13	1.44E-01	9.97E-01	5.15E-22
PLD4	RPL19P12	2.68E-16	4.68E-01	3.49E-01	4.90E-19
PLD6	GHSR	1.07E-11	1.32E-01	4.00E-02	1.48E-28
PLEKHA3	ORMDL2	2.16E-13	4.98E-01	8.43E-01	1.80E-33

Interactions in the MAPK signaling pathway

To show the detailed interaction structure, we presented the results of 331 significant *cis*-trans interactions in the MAPK signaling pathway in the Additional file 2: Table S3 where min *P*-values indicated that the functional regression model with functional response and discrete predictors was used to test for the interaction for all possible pairs of SNPs within two genes and minimum of *P*-values of the tests was listed in the Additional file 2: Table S3. The column "SNP pair" listed their corresponding pair of SNPs reaching the minimum of the *P*-values and their chromosome locations. From Additional file 2: Table S3 we had several significant observations. First, we observed that the majority of interacting genes were located in different chromosomes, which implied that interactions were not caused by the linkage disequilibrium (LD). Second, we observed that large proportions of interacting genes did not show significant evidence of marginal association. This demonstrated that if we only selected the genes with significant marginal association for epistasis analysis, many interactions would be missed. Third, in general, the function-value-based epistasis analysis (BFGM, min *P*-values) had much smaller *P*-values than the summary statistic-based epistasis analysis (SFGM). Fourth, we observed that the genes interacting with the genes in MAPK signaling pathway were in 147 other pathways, including cytokine-cytokine receptor interaction, Cytosolic DNA-sensing pathway, DNA replicationamong others. Fifth, it was interesting to observe that the interacting genes formed a large connected network with 281 nodes and 317 edges (Fig. 6). We observed hub genes *IBA57-AS1* with 67 connections, *HIST1H2AD* with 21 connections, *PRR24* with 18 connections and *ARL6IP4* with 14 connections. *HIST1H2AD* is a core component of nucleosome and plays a central role in transcription regulation. *ARL6IP4* functions as a splicing inhibitor [25].

Gene ontology and KEGG pathway enrichment analysis

Gene ontology enrichment analysis was performed on the genes in the identified 162361 pairs of significant *cis*-trans interactions influencing the transcription to discover overrepresented functional biological groupings with interactions. Our analysis was performed using the biological process, cellular component and molecular function categories of the gene ontology.

Ontology enrichment analysis found that *cis*-trans interactions were significantly enriched in biological

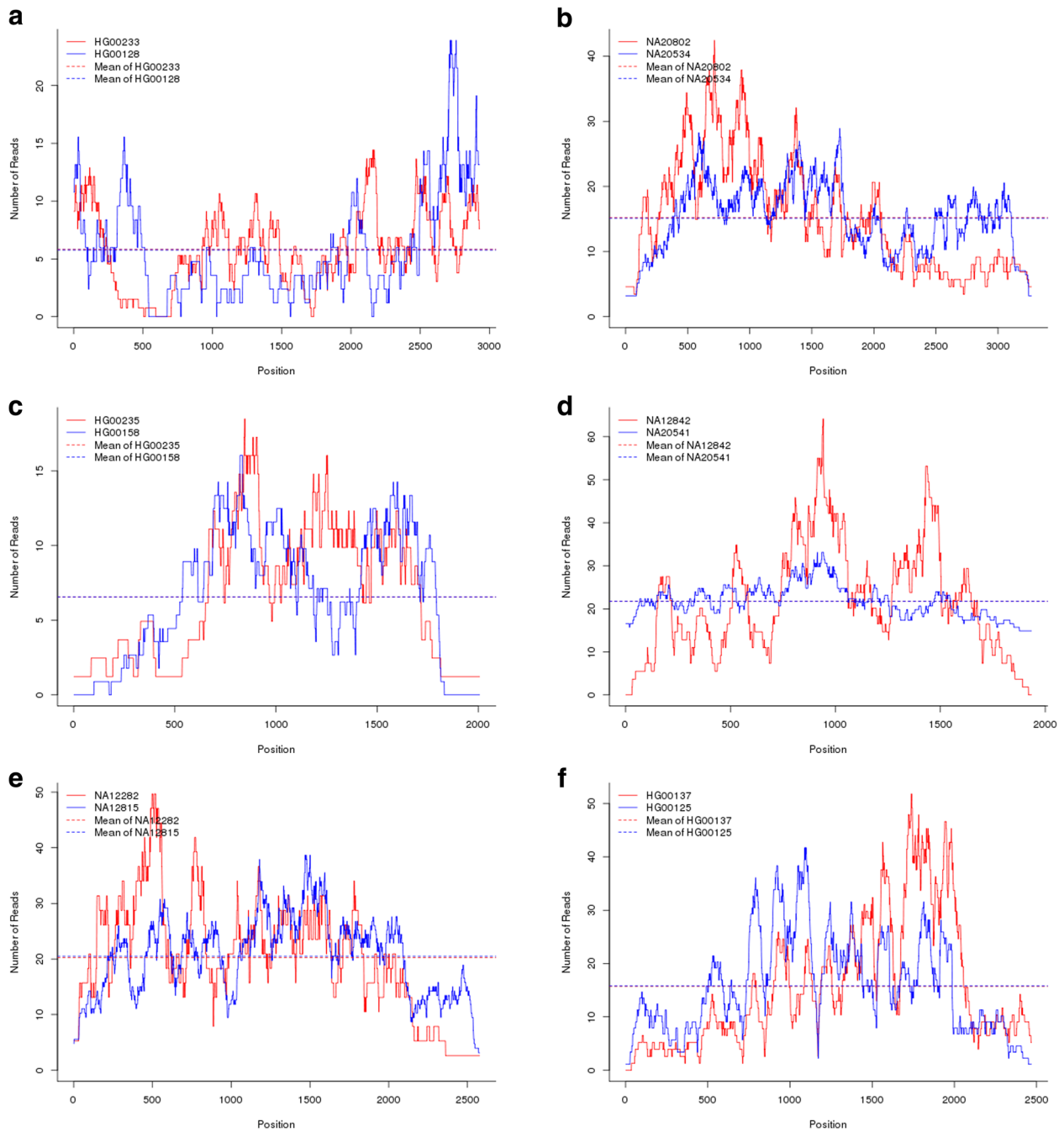

Fig. 5 a. RNA-seq profile of the gene *PLA2G4A* where the curve represented the number of reads as a function of the genomic position. The dotted line denoted the overall expression of the gene *PLA2G4A*. **b**. RNA-seq profile of the gene *PLA2G6* where the curve represented the number of reads as a function of the genomic position. The dotted line denoted the overall expression of the gene *PLA2G6*. **c**. RNA-seq profile of the gene *PLAUR* where the curve represented the number of reads as a function of the genomic position. The dotted line denoted the overall expression of the gene *PLAUR*. **d**. RNA-seq profile of the gene *PLD4* where the curve represented the number of reads as a function of the genomic position. The dotted line denoted the overall expression of the gene *PLD4*. **e**. RNA-seq profile of the gene *PLD6* where the curve represented the number of reads as a function of genomic position. The dotted line denoted the overall expression of the gene *PLD6*. **f**. RNA-seq profile of the gene *PLEKHA3* where the curve represented the number of reads as a function of the genomic position. The dotted line denoted the overall expression of the gene *PLEKHA3*

Table 6 The maximum of r^2 between all possible SNPs in top 20 significantly interacting pairs of genes

GENE1	GENE2	r^2	*P*-value
ULK4	C19orf70	0.00014	0.14974
ULK4	OR10A2	0.00145	0.16383
CCDC13	TMEM121	0.00175	0.19444
ULK4	PSMC5	0.00019	0.11420
ULK4	COX5B	0.00020	0.12871
NKX2-5	TP53TG3D	0.00111	0.18750
ASIC2	RPS16P5	0.00029	0.12406
TMEM132E	LOC100144602	0.00064	0.10577
TMEM98	LOC100144602	0.00043	0.10606
SPACA3	LOC100144602	0.00108	0.09375
ASIC2	OR5B12	0.00048	0.11716
CCL1	TINF2	0.00017	0.13846
SCN2A	DEFB4B	0.00016	0.07823
ZNF254	OR2V1	0.00058	0.14939
KRT5	OR5K1	0.00017	0.03571
FNDC8	LOC100144602	0.00025	0.05882
CCT6B	LOC100144602	0.00029	0.05000
TMEM163	HIST1H4H	0.00193	0.14273
ASIC2	LOC100144602	0.00059	0.12145
KRT5	IFNA7	0.00023	0.07143

processes (BP) including a single organism process, single organism cellular process, single organism metabolic process and development process (Fig. 7) and molecular functions that were primarily related to catalytic and binding activity with *P*-values $<10^{-4}$ (Fig. 8). Ontology enrichment analysis also identified that *cis*-trans interactions were significantly enriched in the cell, intracellular, organelle, and membrane bounded organelle components (Fig. 9).

The enrichment analysis was also applied to 228 KEGG pathways to identify the pathways that were enriched with *cis*-trans interactions. The results were summarized in Fig. 10. The *cis*-trans interactions were enriched in metabolic pathways, MAPK signaling pathway, pathways in cancer, endocytosis, protein processing in endoplasmic reticulum and Wnt signaling pathway.

Communities in gene interaction networks

We used random walks in igraph [26] to detect 10 communities from the entire gene-gene interaction network. We used R package GOstats [27] to conduct gene set enrichment analysis. We have identified 29 pathways enriched in 10 communities. Figure 11 showed the 6^{th} community with 96 genes and 186 interactions enriched with metabolism (Three of four significantly enriched pathways were metabolism pathways: Glycerolipid metabolism, Nicotinate and nicotinamide metabolism, and Pyrimidine metabolism) where node represents a gene and an edge represents the interaction between the connected gene by the edge. All 10 communities with the enriched pathways (*P*-value < 0.01) are summarized in Additional file 3: Table S4.

Discussion

In the past, the statistical epistasis of gene expression is defined as variant–variant interactions that regulate gene expression and its analysis has been mainly designed for microarray gene expression data and common variants. Since the dimension of the data for epistasis analysis of gene expression is very high, all the traditional methods for epistasis analysis of gene expression have the limited application to eQTL data. The whole genome epistasis studies of gene expressions have been very limited. The genetic structure of epistasis of gene expressions has not been fully discovered.

The recently developed next-generation mRNA sequencing (RNA-seq) assay generates dozens or even one hundred million short reads of mRNA and WGS also generates millions of SNPs. As a consequence, these genetic variation and gene expression variation data are so densely distributed across the genome that both genetic variation and expression variation can be modeled as a function of genomic location. The RNA-seq profiles can be taken as a function-valued trait. However, the standard multivariate statistical analysis often fails with functional data. The computational burden and correction for multiple tests seriously damage the feasibility of the variant-variant interaction analysis of extremely high dimensional RNA-seq and WGS genotype data. The variant-variant interaction analysis is not suitable for the epistasis analysis of the function-valued traits with NGS data as genotype data. Although the genetic study of quantitative traits has seen wide application and extensive technical development, the quantitative genetic analysis, particularly epistasis analysis of function-valued trait is comparatively less developed. To our knowledge, no statistical methods have been developed for genetic epistasis analysis of function-valued traits with NGS data. In the past few years we have witnessed the rapid development of novel statistical methods for association studies using NGS data. However, these methods might not be appropriate for genetic epistasis analysis of function-valued trait. The quantitative genetic epistasis analysis of rare variants for function-valued traits remains a huge challenge.

The widely used methods for reducing dimensionality of the RNA-seq data use the Poisson distribution, binomial distribution and negative binomial distribution to summarize the RNA-seq profile into a single number to represent the RNA-seq curve. However, these discrete

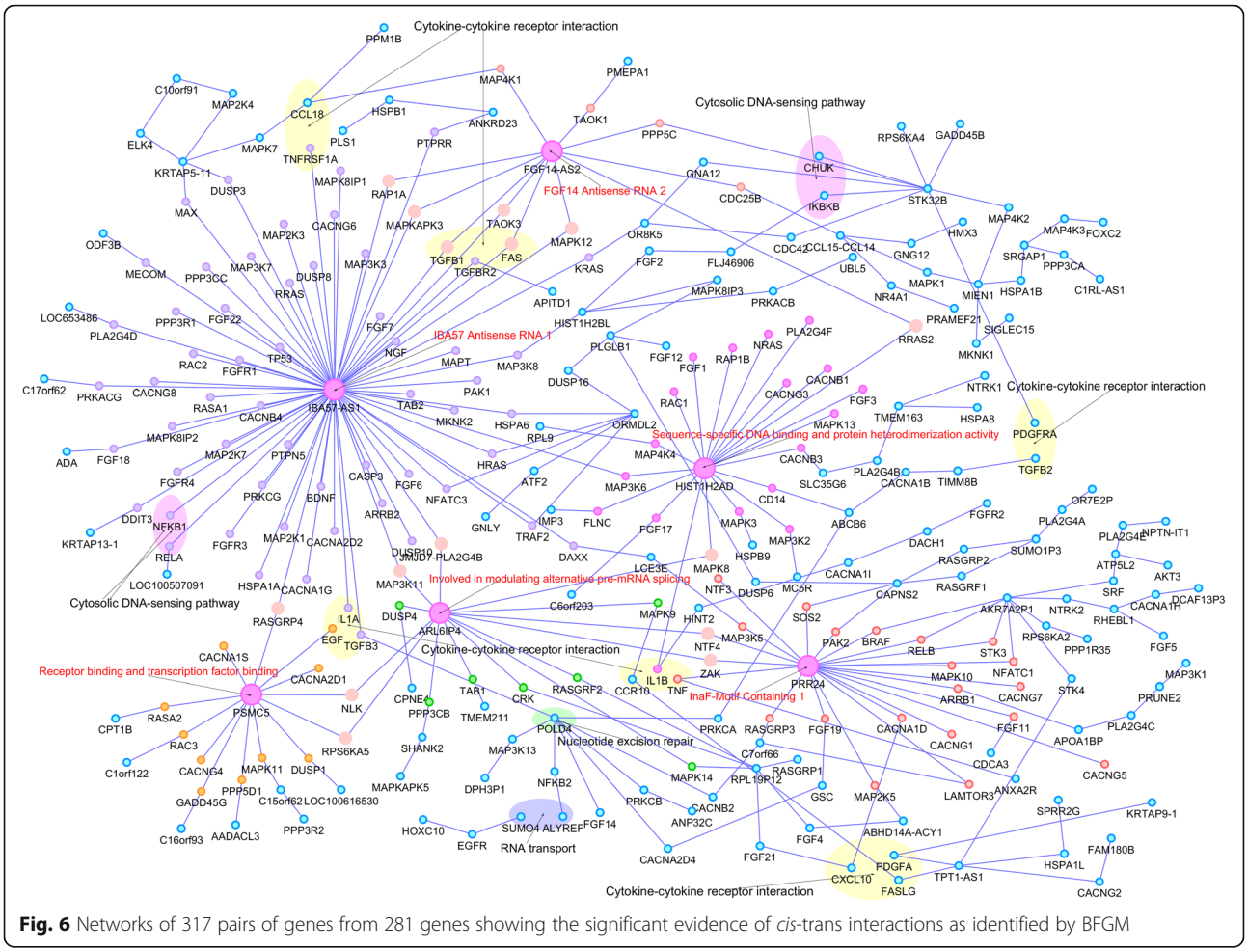

Fig. 6 Networks of 317 pairs of genes from 281 genes showing the significant evidence of *cis*-trans interactions as identified by BFGM

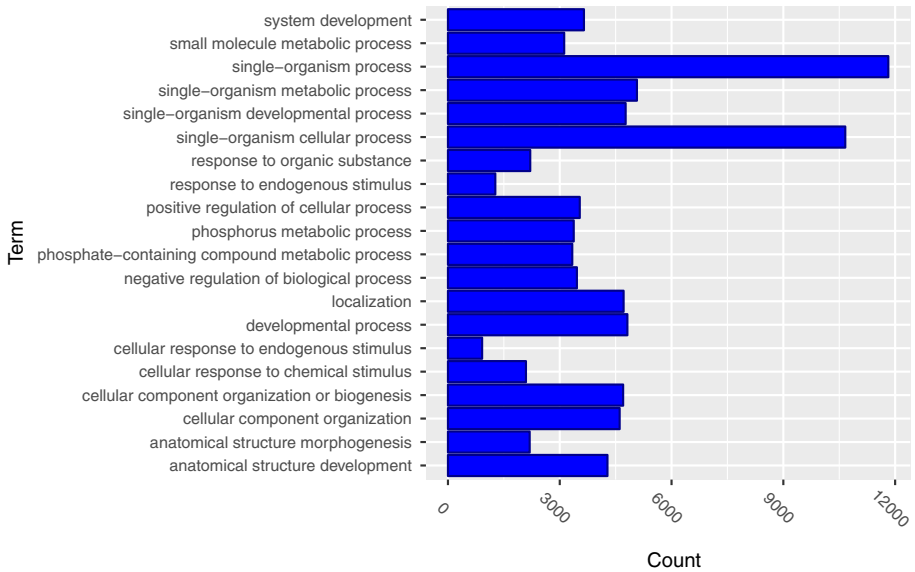

Fig. 7 Gene ontology (GO) enrichment analysis of *cis*-trans interactions: enriched in the category of biological processes

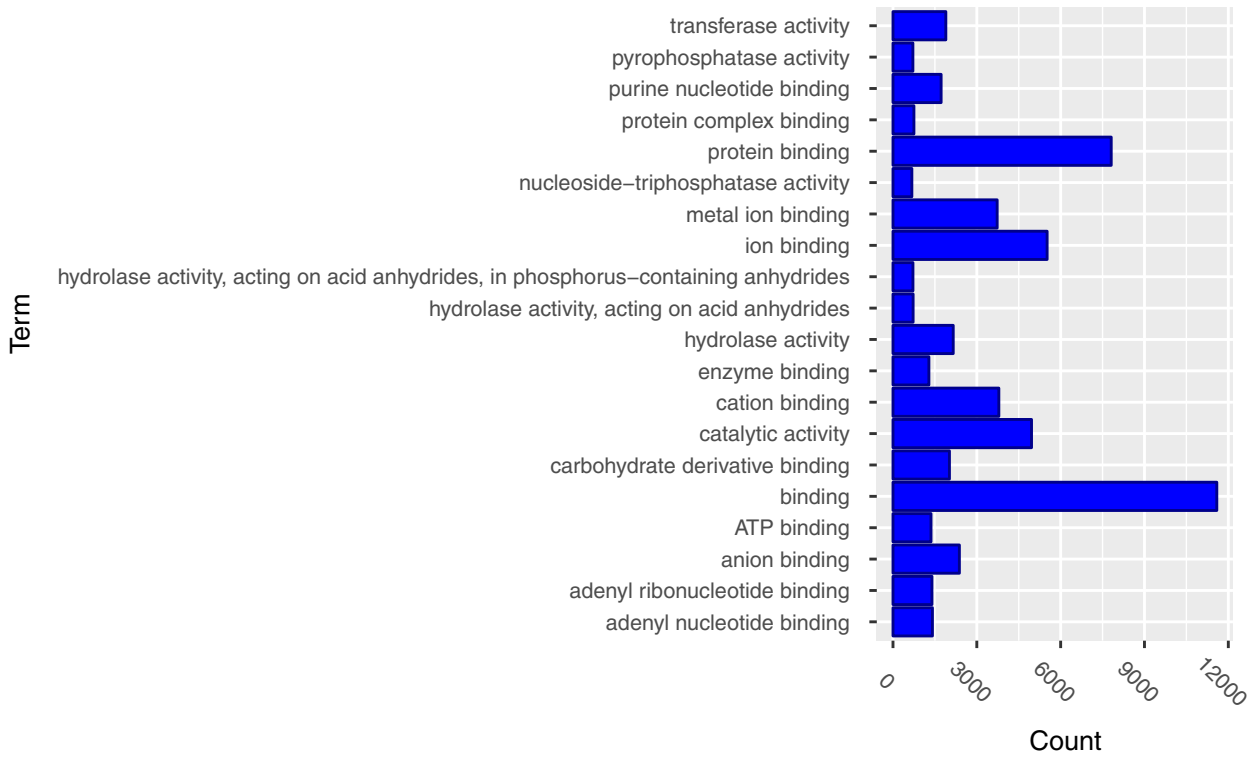

Fig. 8 Gene ontology (GO) enrichment analysis of *cis*-trans interactions: enriched in the category of molecular functions

distributions cannot capture the shape and variation of the RNA-seq curve. To illustrate this we presented Additional file 4: Figure S1A showing the real RNA-seq curve, the data simulated by a negative distribution of the gene *LMNB2* and Additional file 4: Figure S1B showing the real RNA-seq curve of the gene *LMNB2* and the curve estimated by the FPCA of the RNA-seq data. We observed that the negative distribution failed to capture the variation of the RNA-seq profile, but the FPCA approximated the RNA-seq curve exceedingly well.

Emergence of the NGS techniques demands a paradigm shift in the analytic methods for eQTL epistasis

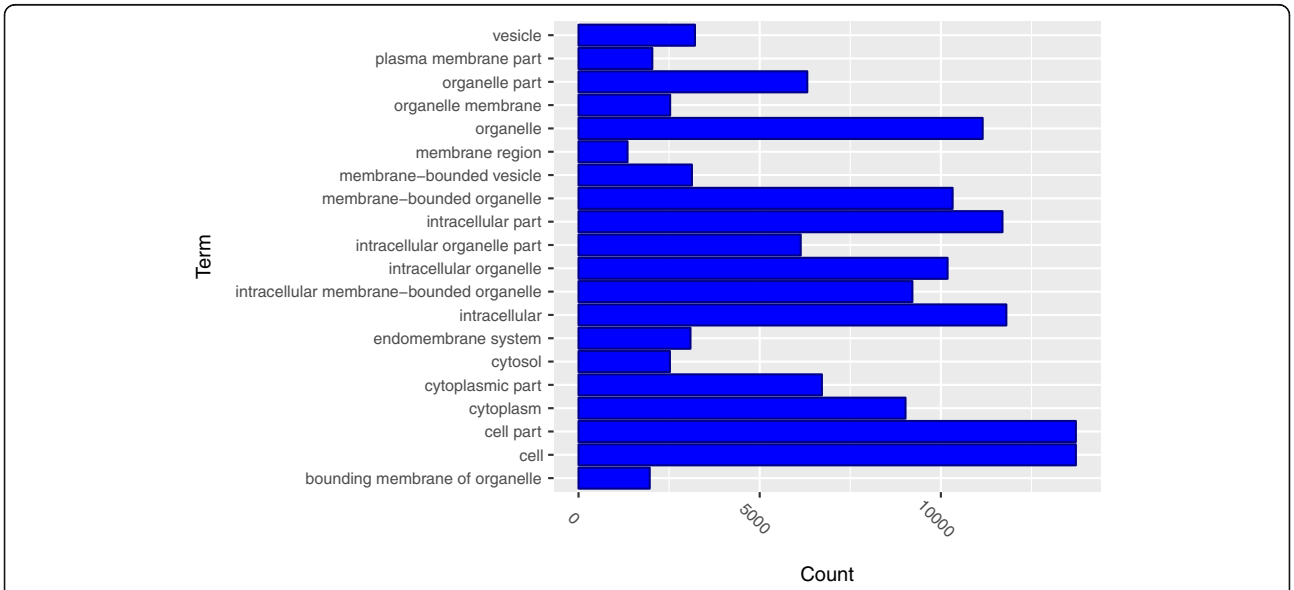

Fig. 9 Gene ontology (GO) enrichment analysis of *cis*-trans interactions: enriched in the category of cellular components

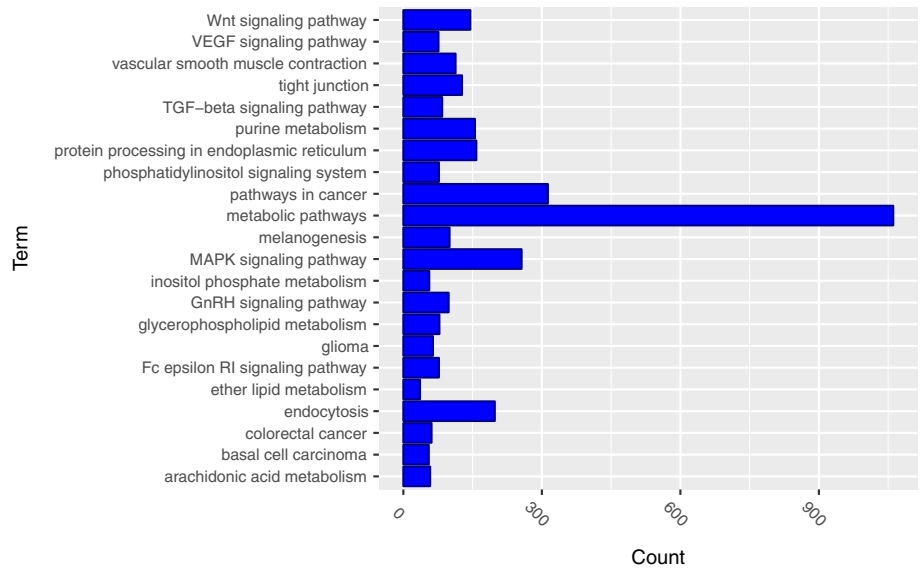

Fig. 10 The enrichment analysis of cis-trans interactions in 228 KEGG pathways

analysis from standard single-variate or multivariate data analysis to functional data analysis. The BFGM with functional response and functional predictors takes a RNA-seq profile as a functional response and genetic variants across the genomic regions as functional predictors, which can be used to test the association of the entire allelic spectrum of the genetic variation with a function-valued trait and has several remarkable features. First, unlike simple and multiple regressions that discard a large amount of information, the BFGM preserves the intrinsic structure and all the positional-level genetic information. Second, the multiple regressions

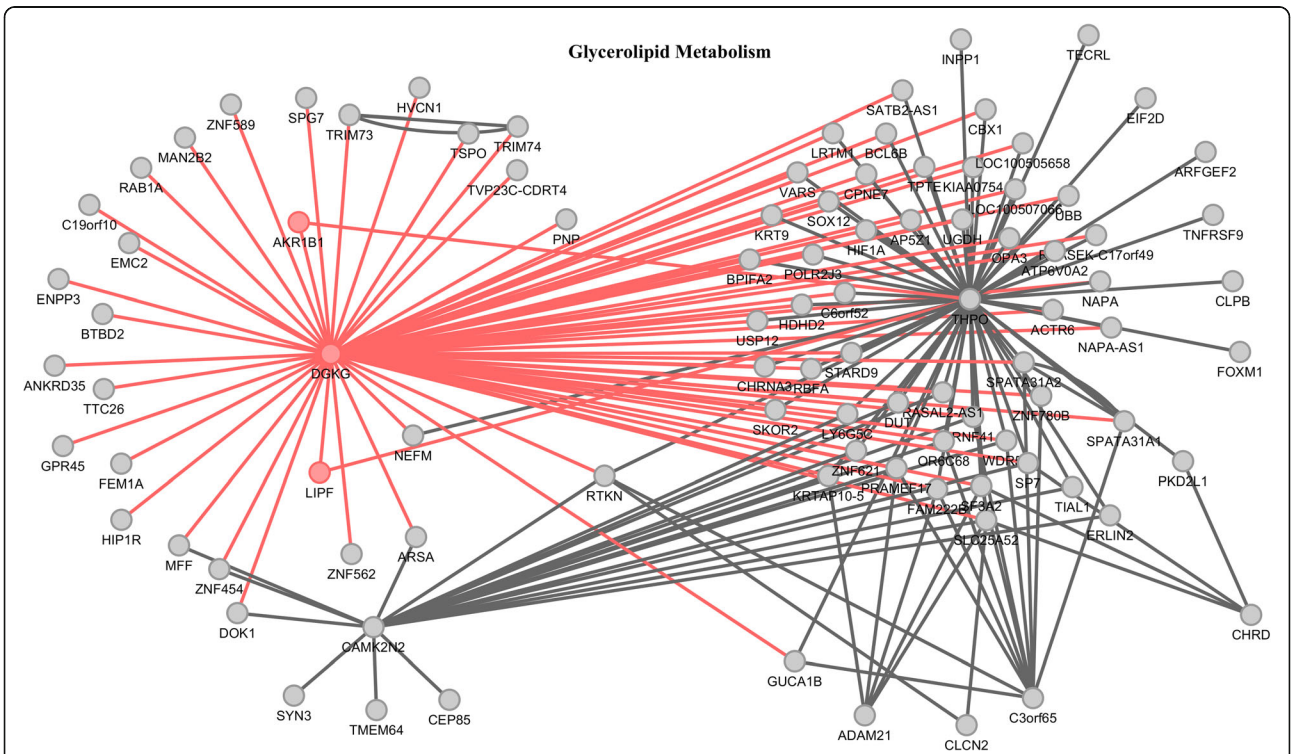

Fig. 11 A subnetwork in the 6^{th} community with 96 genes and 186 interactions enriched with metabolism where node represents a gene and an edge represents the interaction between the connected gene by the edge

will not account for the space-ordering of the data and correlation information contained in the data. The BFGM simultaneously employs genetic information of the individual variants and correlation information contained in both RNA-seq and SNP data. Third, both the sign and the size of the heterogeneity will also be incorporated into the test in the BFGM. Fourth, the multicolinearity problem in the BFGM is alleviated. Fifth, the BFGM expands both RNA-seq function and genotype function in terms of orthogonal eigenfunctions, which leads to substantial dimension reduction. The BFGM for genetic epistasis analysis of a function-valued trait which captures key information in the data is expected to open a new route for genetic epistasis analysis of RNA-seq and NGS genotype data.

Conclusions

We developed a novel functional regression model with both functional response and functional predictors for detection of epistasis influencing RNA-seq variations in humans, which is referred to as the BFGM. The BFGM takes genes as a basic unit of epistasis analysis and utilizes all information contained in both the RNA-seq and SNP data. By large simulations and real data analysis we demonstrated the merits and limitations of the proposed new paradigm of epistasis analysis for the RNA-seq and WGS data.

The new approach uses all genetic information in the genome regions and expression variation information in the target gene to collectively test the interaction between multiple SNPs within the regions influencing the RNA-seq curves. Therefore, the BFGM for interaction analysis overcomes limitations inherent in pair-wise interaction tests with the summary expression level as a scalar trait. By large simulations and real data analysis, we showed that the proposed BFGM substantially increased the power, dramatically reduced the computational burden and substantially outperformed the traditional variant-variant epistasis analysis of summary statistic measured quantitative traits. 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 previous interaction analyses have mainly focused on the interactions between common and common variants. The distribution of the common and rare variants causing interactions is unknown. Very few genome-wide interaction analyses with the RNA-seq and WGS data, and very few results of significant interaction between rare and rare variants, and rare and common variants have been reported. We analyzed 350 samples of European origin with both RNA-seq and whole genome sequencing data available. We observed the large proportions of pair-wise interactions between rare and rare variants, and rare and common variants. The significant pair-wise interactions between common and common variants were less observed. The results showed that the number of significant *cis*-trans interactions identified by the SFGM with RPKM as overall gene expression level only accounted for 0.16% of the significant *cis*-trans interactions identified by the BFGM with RNA-seq and NGS genotype data. The majority of epistasis analysis for gene expressions used the microarray to measure gene expressions and test interactions for only common variants. Even though the RNA-seq data are available they still converted variation rich RNA-seq data into a single number such as RPKM or other summary statistics. Then, the variant-variant epistasis analysis is conducted on these converted data. That explains why these researches question the universe presence of significant gene-gene interaction influencing gene expressions.

Some researchers 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 the majority of the significantly interacting genes showed no marginal association. These results clearly demonstrated that if we tested interactions for only genes with marginal associations, then many true interactions will be missing.

We are unsure whether interaction is most often presented in isolation, or interacting genes form networks. We identified a large number of *cis*-trans interactions and observed that the interacting genes formed large connected networks with hub genes presented. We found that some hub genes, for example histone modification genes, can globally regulate gene expressions. Enrichment analysis also showed that metabolic pathways, MAPK signaling pathway, pathways in cancer, endocytosis, protein processing in endoplasmic reticulum and Wnt signaling pathway among others were enriched with *cis*-trans interactions.

The results in this paper are preliminary. The confounding factors that cause spurious interactions have not been investigated. The statistical methods for epistasis analysis which remove confounding factors have not been developed. The complete genome-wide epistasis analysis including all *cis* and trans interactions have not been performed. The purpose of this paper is to stimulate further discussions regarding the great challenges we are facing in the epistasis analysis of high dimensional RNA-seq and WGS data.

Methods

Functional regression with both functional response and functional predictor models for epistasis analysis

For the convenience of discussion, position level read counts are taken as the RNA-seq profile and is referred

to as a function-valued trait. Let $y_i(\tau)$, $\tau \in T_\tau = [0, T_\tau]$ be the read counts of the i^{th} individual at the genomic position τ. Consider two genomic regions (or genes) $[a_1, b_1]$ and $[a_2, b_2]$. Let $x_i(t)$ and $z_i(s)$ be genotypic functions of the i^{th} individual defined in the regions $[a_1, b_1]$ and $[a_2, b_2]$, respectively. Let t and s be a genomic position in the first and second genomic regions, respectively. The genotype functions $x_i(t)$ and $z_i(s)$ are defined as

$$x_i(t) = \begin{cases} 2P_m(t), & MM \\ P_m(t)\text{-}P_M(t), & Mm, \\ -2P_M(t), & mm \end{cases} \qquad z_i(s) = \begin{cases} 2P_m(s), & MM \\ P_m(s) - P_M(s), & Mm, \\ -2P_M(s), & mm \end{cases}$$

where M and m are two alleles of the marker at the genomic position t and s, $P_M(t)$ and $P_m(t)$, and $P_M(s)$, $P_m(s)$ are the frequencies of the alleles M and m at the genomic positions t and s, respectively. Consider a functional regression model with functional response and functional predictors (BFGM):

$$y_i(\tau) = \mu(\tau) + W_i^T\omega(\tau) + \int_T x_i(t)\alpha(t,\tau)dt + \int_S z_i(s)\beta(s,\tau)ds + \int_T\int_S x_i(t)z_i(s)\gamma(t,s,\tau)dsdt + \varepsilon_i(\tau) \tag{1}$$

where $\mu(\tau)$ is an overall mean function at the genomic position τ, W_i is a vector of covariates for i^{th} individual, $\omega(\tau)$ is a vector of effects associated with the covariates, $\alpha(t, \tau)$ is a genetic additive effect function at genomic position t of the first gene and genomic position τ of the RNA-seq profile, $\beta(s, \tau)$ is a genetic additive effect function at genomic positions s of the second gene and the genomic position τ, $\gamma(t, s, \tau)$ is an interaction effect function between two putative quantitative trait loci (QTLs) located at the genomic positions t and s influencing the read counts at the genomic position τ, and $\varepsilon_i(\tau)$ is a residual function of the unexplained effect for the i^{th} individual at the genomic position τ. The interaction function is measured by double integrals of the genotype function in two genes.

Estimation of interaction effect function

We assume that both position level read count function and genotype functions are centered. The genotype functions $x_i(t)$ and $z_i(s)$ are expanded in terms of the orthonormal basis function as:

$$x_i(t) = \sum_{j=1}^{\infty}\xi_{ij}\phi_j(t) \quad \text{and} \quad z_i(s) = \sum_{l=1}^{\infty}\eta_{il}\psi_l(s), \tag{2}$$

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

$$\xi_{ij} = \int_T x_i(t)\phi_j(t)dt \quad \text{and} \quad \eta_{il} = \int_S z_i(s)\psi_l(s)ds \tag{3}$$

In practice, numerical methods for the integral will be used to calculate the expansion coefficients. Substituting Eq. (2) into Eq. (1), we obtain

$$y_i(\tau) = \mu(\tau) + W_i^T\omega(\tau) + \sum_{j=1}^{J}\xi_{ij}\alpha_j(\tau) + \sum_{l=1}^{L}\eta_{il}\beta_l(\tau) + \sum_{j=1}^{J}\sum_{l=1}^{L}\xi_{ij}\eta_{il}\gamma_{jl}(\tau) + \varepsilon_i(\tau), \tag{4}$$

where

$$\alpha_j(\tau) = \int_T \alpha(t,\tau)\phi_j(t)dt,\ \beta_l(\tau) = \int_S \beta(s,\tau)\psi_l(s)ds \text{ and}$$

$$\gamma_{jl}(\tau) = \int_T\int_S \gamma(t,s,\tau)\phi_j(t)\psi_l(s)dtds.$$

The parameters $\alpha_j(\tau)$, $\beta_l(\tau)$ and $\gamma_{jl}(\tau)$ are referred to as genetic additive effect and additive x additive effect score functions. These score functions can also be viewed as the expansion coefficients of the genetic effect functions with respect to orthonormal basis functions:

$$\begin{aligned}\alpha(t,\tau) &= \sum_j \alpha_j(\tau)\phi_j(t), \beta(s,\tau) \\ &= \sum_l \beta_l(\tau)\psi_l(s) \text{ and } \gamma(s,t) \\ &= \sum_j\sum_l \gamma_{jl}(\tau)\phi_j(s)\psi_l(t)\end{aligned}$$

Equation (4) can be written in a vector form:

$$Y(\tau) = E\mu(\tau) + W\omega(\tau) + \xi\alpha(\tau) + \eta\beta(\tau) + \Gamma\gamma(\tau) + \varepsilon(\tau)\ , \tag{5}$$

where $Y(\tau)$, $\mu(\tau)$, $\omega(\tau)$, $\alpha(\tau)$, $\beta(\tau)$ and $\gamma(\tau)$ are vectors, W, ξ, η and Γ are matrices.

Expanding $Y(\tau)$, $\mu(\tau)$, $\omega(\tau)$, $\alpha(\tau)$, $\beta(\tau)$, $\gamma(\tau)$ and $\varepsilon(\tau)$ in terms of the orthogonal basis functions yield

$$y_i(\tau) = \sum_{k=1}^{K} y_{ik}\theta_k(\tau)\ ,\ \mu(\tau) = \sum_{k=1}^{K}\mu_k\theta_k(\tau)\ ,$$

$$\omega_j(\tau) = \sum_{k=1}^{K}\omega_{jk}\theta_k(\tau)\ ,\ \alpha_j(\tau) = \sum_{k=1}^{K}\alpha_{jk}\theta_k(\tau)\ ,$$

$$\beta_j(\tau) = \sum_{k=1}^{K}\beta_{jk}\theta_k(\tau)\ ,\ \gamma_{jl}(\tau) = \sum_{k=1}^{K}\gamma_{jlk}\theta_k(\tau)\ ,$$

$$\text{and}\quad \varepsilon_i(\tau) = \sum_{k=1}^{K}\varepsilon_{ik}\theta_k(\tau).$$

Define expansion coefficient vectors and matrices as follows.

$$Y=\begin{bmatrix} y_{11} & \cdots & y_{1K} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nK} \end{bmatrix},\ \mu=\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_K \end{bmatrix}^{\mathrm{T}},\ E=\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\omega=\begin{bmatrix} \omega_{11} & \cdots & \omega_{1K} \\ \vdots & \ddots & \vdots \\ \omega_{d1} & \cdots & \omega_{dK} \end{bmatrix},\ \alpha=\begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1K} \\ \vdots & \ddots & \vdots \\ \alpha_{J1} & \cdots & \alpha_{JK} \end{bmatrix},$$

$$\beta=\begin{bmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{L1} & \cdots & \beta_{LK} \end{bmatrix},\ \gamma=\begin{bmatrix} \gamma_{111} & \cdots & \gamma_{11K} \\ \vdots & \ddots & \vdots \\ \gamma_{JL1} & \cdots & \gamma_{JLK} \end{bmatrix}$$

and $\varepsilon=\begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1K} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nK} \end{bmatrix}$.

Thus, substituting the above expansion into Eq. (5) gives

$$\begin{aligned} Y\theta(\tau)=\mu\theta(\tau)+W\omega\theta(\tau)+\xi\alpha\theta(\tau)+\eta\beta\theta(\tau) \\ +\Gamma\gamma\theta(\tau)+\varepsilon\theta(\tau), \end{aligned} \tag{6}$$

where

$$W=\begin{bmatrix} W_{11} & \cdots & W_{1d} \\ \vdots & \ddots & \vdots \\ W_{n1} & \cdots & W_{nd} \end{bmatrix},\quad \xi=\begin{bmatrix} \xi_{11} & \cdots & \xi_{1J} \\ \vdots & \ddots & \vdots \\ \xi_{n1} & \cdots & \xi_{nJ} \end{bmatrix},$$

$$\eta=\begin{bmatrix} \eta_{11} & \cdots & \eta_{1L} \\ \vdots & \ddots & \vdots \\ \eta_{n1} & \cdots & \eta_{nL} \end{bmatrix}\ \text{and}\ \Gamma=\begin{bmatrix} \xi_1^T\otimes\eta_1^T \\ \vdots \\ \xi_n^T\otimes\eta_n^T \end{bmatrix}$$

$$=\begin{bmatrix} \xi_{11}\eta_{11} & \cdots & \xi_{11}\eta_{1L} & \cdots & \xi_{1J}\eta_{11} & \cdots & \xi_{1J}\eta_{1L} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \xi_{n1}\eta_{n1} & \cdots & \xi_{n1}\eta_{nL} & \cdots & \xi_{nJ}\eta_{n1} & \cdots & \xi_{nJ}\eta_{nL} \end{bmatrix}.$$

Since Eq. (6) holds for every genomic position τ, the coefficients on both sides of Eq. (6) should be equal. Therefore, the functional regression model (6) can be further transformed to standard multivariate multiple regression:

$$Y=E\mu+W\omega+\xi\alpha+\eta\beta+\Gamma\gamma+\varepsilon. \tag{7}$$

Let

$$A=[\,E\quad W\quad \xi\quad \eta\quad \Gamma\,]\ \text{and}\ b=\begin{bmatrix} \mu \\ \omega \\ \alpha \\ \beta \\ \gamma \end{bmatrix}.$$

Equation (7) can be rewritten as

$$Y=Ab+\varepsilon. \tag{8}$$

The standard least square estimators of b is

$$\hat{b}=\left(A^TA\right)^{-1}A^TY. \tag{9}$$

The covariance matrix Σ is estimated by

$$\hat{\Sigma}=\frac{\left(Y-A\hat{b}\right)^T\left(Y-A\hat{b}\right)}{(n-(1+d+J+L+JL))\mathrm{K}}. \tag{10}$$

Test statistic

An essential problem in genetic epistasis analysis of the function-valued traits is to test the interaction between two genomic regions (or genes). Formally, we investigate the problem of testing the following hypothesis:

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

which is equivalent to testing the hypothesis:

$$H_0:\gamma=0. \tag{11}$$

Let *vec* denote the vector operation. To develop test statistics, we begin with calculating the covariance matrix of the $vec\left(\hat{b}\right)$. We assume that

$$\mathrm{var}(vec(\varepsilon))=\Sigma\otimes I_n. \tag{12}$$

Recall that

$$vec\left(\hat{b}\right)=\left[I_K\otimes\left(A^TA\right)^{-1}A^T\right]vec(Y).$$

Therefore, we have

$$\begin{aligned} \mathrm{var}\left(vec\left(\hat{b}\right)\right)&=\left[I_K\otimes\left(A^TA\right)^{-1}A^T\right](\Sigma\otimes I_n)\left[I_K\otimes A\left(A^TA\right)^{-1}\right] \\ &=\Sigma\otimes\left(A^TA\right)^{-1} \end{aligned} \tag{13}$$

Let Λ be a matrix consisting of the last JLK columns and JLK rows of the covariance matrix $\mathrm{var}\left(vec\left(\hat{b}\right)\right)$ and $\hat{\gamma}$ be the estimators of interaction which can be obtained by extracting the last JL rows of the estimators of the matrix $\hat{b}$. Define the test statistic for testing the interaction between two genomic regions $[a_1,b_1]$ and $[a_2,b_2]$ as

$$T_I=vec(\hat{\gamma})^T\Lambda^{-1}vec(\hat{\gamma}). \tag{14}$$

Then, under the null hypothesis $H_0:\gamma=0$, T_I is asymptotically distributed as a central $\chi^2_{(JLK)}$ with degrees of freedom JLK or the rank of the matrix Λ.

Null distribution of test statistics

To examine the null distribution of test statistics, we performed a series of simulation studies to compare their empirical levels with the nominal ones. We calculated the type I error under three models. We first assumed the model with no marginal effects:

Model 1 (no marginal effect):

$$y_i(\tau) = \mu(\tau) + \varepsilon_i(\tau), \tag{15}$$

where $\mu(\tau)$ is the overall mean at the genomic position τ, $y_i(\tau)$ is the normalized number of reads at the genomic position τ of the i^{th} individual and $\varepsilon_i(\tau)$ is an error stochastic process. The errors should be correlated stochastic process. The theoretic models for the errors are unclear. They were estimated from the data. The procedures for generating mean $\mu(\tau)$ and errors $\varepsilon_i(\tau)$ consisted of the following steps.

Step 1: We randomly sampled 100 genes from the whole real RNA-seq dataset. Let k index genes, j index the genomic positions and i index the samples. Assume that the gene k is located in the interval $[a_k, b_k]$. Let x_{ikj} , $(i = 1, ..., n, k = 1, ..., 100, j = 1, ..., s_k)$ be the observed count of reads of the gene k in the genomic position j of the i^{th} individual where the length of gene k is denoted s_k. For each genomic position, we define an n dimensional vector:

$x_{kj} = [x_{1kj}, ..., x_{nkj}]^T$.

Step 2. Let m be the median length of 100 genes. In our dataset, $m = 2,456$.

Step 3. Re-map the original RNA-seq data of 100 genes to the interval [0, 1] using transformation $\frac{j-a_k}{b_k-a_k}$. Then, estimate the count of reads on position $0, \frac{1}{m}, \frac{2}{m}, ..., 1$ from the original RNA-seq data of the 100 genes using local polynomial regression (LOESS). The estimated count of reads of the gene k in the genomic position j of the i^{th} individual at the equally distributed new positions $0, \frac{1}{m}, \frac{2}{m}, ..., 1$ are denoted by y_{ikj}. Define vector

$$y_{kj} = \left[y_{1kj}, ..., y_{nkj}\right]^T, k = 1, ..., 100, j = 1, ..., m.$$

Step 4. Compute the means of the re-mapped the RNA-seq data over 100 genes and over n samples: $\overline{y}_{ij} = \frac{1}{100}\sum_{k=1}^{100} y_{ikj}, i = 1, ..., n, j = 1, ..., m$ and $\overline{y}_j = \frac{1}{100\times n}\sum_{i=1}^{n}\sum_{k=1}^{100} y_{ikj}, j = 1, ..., m.$

Define the mean vector of the re-mapped counts of reads: $\overline{y} = [\overline{y}_1, ..., \overline{y}_m]^T$.

Step 5. Compute the mean function $\mu(\tau)$. Pooling all the re-mapped data:

$$Y = \begin{bmatrix} \overline{y}_{11} \cdots \overline{y}_{1m} \\ \vdots \vdots \vdots \\ \overline{y}_{n1} \vdots \overline{y}_{nm} \end{bmatrix}.$$

Use the pooled data to perform FPCA, which leads to functional principal component expansion:

$$y_i(\tau) = \sum_{l=1}^{L} \xi_{il}\beta_l(\tau), i = 1, ..., n,$$

where $\beta_l(\tau)$ are the functional principal components.

Calculate $\overline{\xi}_l = \frac{1}{n}\sum_{i=1}^{n} \xi_{il}, l = 1, ..., L$. Using the averaged functional principal component score, we compute the mean $\mu(\tau)$ as follows:

$$\mu(\tau) = \sum_{l=1}^{L} \overline{\xi}_l\beta_l(\tau).$$

Step 6. Define the centralized RNA-seq data matrix:

$$Z = \begin{bmatrix} z_{11} \cdots z_{1m} \\ \vdots \vdots \vdots \\ z_{n1} \cdots z_{nm} \end{bmatrix},$$

where $Z_{ij} = \overline{y}_{ij} - \overline{y}_j, i = 1, ..., n, j = 1, ..., m$.

We perform FPCA on the centralized dataset Z where means of the RNA-seq data at each genomic position over 100 genes is removed and obtain a set of functional principal components (eigenfunctions) $\{\phi_1(\tau), ..., \phi_T(\tau)\}$ and functional principal component scores $\eta_{it}, i = 1, ..., n, t = 1, ..., T$. Define T random variables $\eta = [\eta_1, ..., \eta_T]$ with vectors of their sampling values:$\eta_t = [\eta_{1t}, ..., \eta_{nt}]^T$, $t = 1, ..., T$.

We then calculate the sampling covariance matrix $\hat{\Sigma} = \mathrm{cov}(\eta, \eta)$.Assume that the scores of the residuals follow a multivariate normal distribution $N(0, \hat{\Sigma})$. Using the normal random variables to generate an n sample of vectors $\varepsilon_i = [\varepsilon_{i1}, ..., \varepsilon_{iT}]$. The residuals $\varepsilon_i(\tau)$ will be defined as

$$\varepsilon_i(\tau) = \sum_{t=1}^{T} \varepsilon_{it}\phi_t(\tau), i = 1, ..., n.$$

Model 2 (a marginal effect at the first gene):

$$y_i(\tau) = \mu(\tau) + \sum_{j=1}^{J} x_{ij}\alpha_j(\tau) + \varepsilon_i(\tau), \tag{16}$$

where $\mu(\tau)$ is the overall mean at the genomic position τ, $\varepsilon_i(\tau)$ is an error stochastic process, x_{ij} is an indicator variable for the genotype of i^{th} individual at the j^{th} SNP of the first gene, $y_i(\tau)$ is defined as that in model 1. The coefficient $\alpha_j(\tau) = r_j \cdot \alpha(\tau)$, where r_j is randomly selected from 0.5 to 1.5, is the additive effect function of the j^{th} SNP of the first gene, $\mu(\tau)$ is obtained by randomly sampling 100 genes from the real RNA-seq and WGS genotype data without interactions and $\alpha(\tau)$ is obtained by randomly sampling 100 genes from the real RNA-seq and

WGS genotype data under the condition that one gene have significant main effect, the other gene do not have significant main effect, and these gene pairs are not in the list of significantly interacted gene pairs in our results. The overall mean $\mu(\tau)$, effect function $\alpha(\tau)$ and the residuals $\varepsilon_i(\tau)$ were similarly simulated as that in Model 1.

Model 3 (marginal effects at both the first and the second genes):

$$y_i(\tau) = \mu(t) + \sum_{j=1}^{J} x_{ij}\alpha_j(\tau) + \sum_{k=1}^{K} z_{ik}\beta_k(\tau) + \varepsilon_i(\tau), \quad (17)$$

where z_{ik} is an indicator variable for the genotype of i^{th} individual at the k^{th} SNP of the second gene. The genetic additive effect function $\beta_k(\tau_l)$ is assumed to be equal to $\beta_k(\tau) = s_k\beta(\tau)$, where s_k is randomly selected from 0.5 to 1.5, other parameters are defined in Model 2. A total of 100 pairs of genes were randomly selected under the condition that both genes have significant main effect and these gene pairs are not in the list of significant interacted gene pairs in our results. The overall mean function $\mu(\tau)$, main effect functions $\alpha(\tau)$ and $\beta(\tau)$, and residual term $\varepsilon_i(\tau)$ were similarly generated as that in Models 1 and 2.

Power evaluation

To evaluate the performance of the functional regression model for testing epistatic effects on gene expression, we estimated the power through simulations. We assumed that there were k_1 SNPs in the first gene and k_2 SNPs in the second gene. Thus, there were totally k_1k_2 SNP pairs between these two genomic regions. For the h^{th} pair of SNPs, let Q_{h_1} and q_{h_1} be two alleles at the SNP in the first gene, Q_{h_2} and q_{h_2} be two alleles at the SNP in the second gene. Let u^h_{ijkl} denote her/his genotypes of the h^{th} pair of SNPs, where $ij \in Q_{h_1}Q_{h_1}, Q_{h_1}q_{h_1}, q_{h_1}q_{h_1}$ and $kl \in Q_{h_2}Q_{h_2}, Q_{h_2}q_{h_2}, q_{h_2}q_{h_2}$. Let $g^h_{u_{ijkl}}(\tau)$ denote her/his genotypic value in the h^{th} pair of SNPs at genomic position τ influencing gene expressions. Then we can use the following multiple regression model to generate the function-valued trait (RNA-seq) of the u^{th} individual of the h^{th} pair of SNPs at the genomic position τ.

$$y_u(\tau) = \sum_{h=1}^{k_1k_2} g^h_{u_{ijkl}}(\tau) + \varepsilon_u(\tau), u = 1, \ldots, n, \quad (18)$$

$g^h_{u_{ijkl}}(\tau) = \lambda^h_{u_{ijkl}}\, g(\tau)$, $\lambda^h_{u_{ijkl}}$ is a risk parameter which is determined by the gene interaction model (Additional file 5: Table S2), the risk parameter r varies from 0 to 1, and $g(\tau)$ is a common genotype coefficient function fitted by the real RNA-seq data and $\varepsilon_u(\tau)$ is the error stochastic process and estimated from the null model as that in null distribution in the test statistics section.

Additional files

Additional file 1: Table S1. Summary statistics of genetic variants in the five genes. (XLS 25 kb)

Additional file 2: Table S3. A list of significant cis-trans interactions in the MAPK pathway. (XLS 114 kb)

Additional file 3: Table S4. Ten communities and their significantly enrished pathways. (XLSX 12 kb)

Additional file 4: Figure S1. A. The real RNA-seq curve of the gene *LMNB2* and the data simulated by the negative distribution of the gene *LMNB2*. B. The real RNA-seq curve of the gene *LMNB2* and the curve estimated by the FPCA of the RNA-seq data. (ZIP 120 kb)

Additional file 5: Table S2. The value of risk parameter in different interaction models. (XLS 30 kb)

Abbreviations

BFGM: FRGM with Functional Response and Functional Predictors; FRGM: Functional regression model; NGS: Next generation sequencing; PCA: Principal component analysis; SFGM: FRGM with scalar response and functional predictors; WGS: Whole genome sequencing

Acknowledgement

The authors express appreciation to the computation sources from the Texas Advanced Computing Center and the High-End Computing Center at Fudan University.

Funding

Mr. Xiong was supported by Grant 1R01AR057120–01 and 1R01HL106034-01, from the National Institutes of Health and NHLBI. Ms. Xu and Mr. Jin were supported by research grants from the National Natural Science Foundation of China (31330038, 31521003), National Basic Research Program (2014CB541801), and 111 Project (B13016). The funders had no role in the design of the study, data collection, analysis, or manuscript preparation.

Availability of data and materials

The R package FRGM_1.0.tar.gz is available from https://sph.uth.edu/research/centers/hgc/xiong/software.htm. All results represented in this paper could be reproduced by the examples and programs in this package. SNP data supporting the conclusions of this article are available from FTP site hosted at the EBI, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, and RNA-seq data are available in Geuvadis RNA sequencing project of 1000 Genomes samples under accession E-GEUV-3, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-3/processed/.

Authors' contributions

KLX, JL and MMX designed this study, developed the statistical methods and wrote the manuscript. KLX wrote the R package and performed data analysis. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

This study is a retrospective analysis of the existing data. SNP data are downloaded from FTP site hosted at the EBI, and RNA-seq data are downloaded from EBI ArrayExpress.

Publisher's Note

Author details

[1]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University,

Shanghai 200438, China. [2]School of Data Science and Institute for Big Data, Fudan University, Shanghai 200433, China. [3]Department of Biostatistics, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [4]Human Genetics Center, The University of Texas Health Science Center at Houston, P.O. Box 20186, Houston, TX 77225, USA.

Received: 21 November 2016 Accepted: 9 May 2017
Published online: 18 May 2017

References

1. Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. Trans Roy Soc Edinb. 1918;52:399–433.
2. Lehner B. Molecular mechanisms of epistasis within and between genes. Trends Genet. 2011;27:323–31.
3. Phillips PC. The language of gene interaction. Genetics. 1998;149:1167–71.
4. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9:855–67.
5. Hemani G, Shakhbazov K, Westra H-J, Esko T, Henders AK, McRae AF, Yang J, Gibson G, Martin NG, Metspalu A. Detection and replication of epistasis influencing transcription in humans. Nature. 2014;508:249–53.
6. Huang Y, Wuchty S, Przytycka TM. eQTL epistasis–challenges and computational approaches. Front Genet. 2013;4:51.
7. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci. 2012;109:1193–8.
8. Kang H, Yang X, Chen R, Zhang B, Corona E, Schadt E, Butte A. Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and coexpression networks reveal novel candidate genes for type 2 diabetes. Diabetologia. 2012;55:2205–13.
9. Shang J, Zhang J, Sun Y, Liu D, Ye D, Yin Y. Performance analysis of novel methods for detecting epistasis. BMC Bioinf. 2011;12:1.
10. Kang M, Zhang C, Chun H-W, Ding C, Liu C, Gao J. eQTL epistasis: detecting epistatic effects and inferring hierarchical relationships of genes in biological pathways. Bioinformatics. 2015;31:656–64.
11. Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. Epistatic selection between coding and regulatory variation in human evolution and disease. Am J Hum Genet. 2011;89:459–63.
12. Sun X, Lu Q, Mukherjee S, Crane PK, Elston R, Ritchie MD. Analysis pipeline for the epistasis search–statistical versus biological filtering. Front Genet. 2014;5:106.
13. Lee J, Ji Y, Liang S, Cai G, Müller P. On differential gene expression using RNA-seq data. Cancer Informat. 2011;10:205–15.
14. Li JJ, Jiang C-R, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci. 2011;108:19867–72.
15. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12:671–82.
16. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
17. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct Genomics. 2015;14:130–42.
18. Gosik K, Kong L, Chinchilli VM, Wu R. iFORM/eQTL: an ultrahigh-dimensional platform for inferring the global genetic architecture of gene transcripts. Brief Bioinform. 2017;18(2):250–9.
19. Zhang F, Boerwinkle E, Xiong M. Epistasis analysis for quantitative traits by functional regression model. Genome Res. 2014;24:989–98.
20. Zhang F, Xie D, Liang M, Xiong M. Functional Regression Models for Epistasis Analysis of Multiple Quantitative Traits. PLoS Genet. 2016;12:e1005965.
21. Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
22. Graffelman J, Moreno V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat Appl Genet Mol Biol. 2013;12:433–48.
23. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:1.
24. Beasley TM, Erickson S, Allison DB. Rank-based inverse normal transformations are increasingly used, but are they merited? Behav Genet. 2009;39:580–95.
25. Li Q, Zhao H, Jiang L, Che Y, Dong C, Wang L, Wang J, Liu L. An SR-protein induced by HSVI binding to cells functioning as a splicing inhibitor of viral pre-mRNA. J Mol Biol. 2002;316:887–94.
26. Csardi G, Nepusz T. The igraph software package for complex network research. Inter J Complex Systems. 2006;1695(5):1–9.
27. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23(2):257–8.