

Biological Knowledge-Driven Analysis of Epistasis in Human GWAS with Application to Lipid Traits

Li Ma, Alon Keinan, and Andrew G. Clark

Abstract

While the importance of epistasis is well established, specific gene–gene interactions have rarely been identified in human genome-wide association studies (GWAS), mainly due to low power associated with such interaction tests. In this chapter, we integrate biological knowledge and human GWAS data to reveal epistatic interactions underlying quantitative lipid traits, which are major risk factors for coronary artery disease. To increase power to detect interactions, we only tested pairs of SNPs filtered by prior biological knowledge, including GWAS results, protein–protein interactions (PPIs), and pathway information. Using published GWAS and 9,713 European Americans (EA) from the Atherosclerosis Risk in Communities (ARIC) study, we identified an interaction between *HMGCR* and *LIPC* affecting high-density lipoprotein cholesterol (HDL-C) levels. We then validated this interaction in additional multiethnic cohorts from ARIC, the Framingham Heart Study, and the Multi-Ethnic Study of Atherosclerosis. Both *HMGCR* and *LIPC* are involved in the metabolism of lipids and lipoproteins, and *LIPC* itself has been marginally associated with HDL-C. Furthermore, no significant interaction was detected using PPI and pathway information, mainly due to the stringent significance level required after correcting for the large number of tests conducted. These results suggest the potential of biological knowledge-driven approaches to detect epistatic interactions in human GWAS, which may hold the key to exploring the role gene–gene interactions play in connecting genotypes and complex phenotypes in future GWAS.

Key words Epistasis, Gene–Gene Interaction, Biological Knowledge, GWAS, Lipid

1 Introduction

As of June 2013, over 1,638 publications and 10,859 single nucleotide polymorphisms (SNPs) have been included in the catalog of human genome-wide association studies (GWAS) [1]. Although these SNPs are significantly associated with human diseases and traits, most of them have small effects that, in aggregate, account for only a moderate proportion of heritable variance [2–6]. Four lipid traits, total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), triglyceride (TG), and high-density lipoprotein cholesterol (HDL-C) levels, are among the most important risk factors for coronary heart disease. Recently, meta-analyses of many

GWAS, with a combined sample size of up to 100,000, have detected hundreds of loci associated with levels of these four lipids [7, 8]. However, these loci collectively only explain 25–30 % of heritable variance of each lipid [7, 8]. Many hypotheses have been offered to explain the missing heritability, including rare and structural variants, gene–environment interactions, epigenetics, and complex inheritance [2–5]. The missing heritability may also be partially attributed to epistatic or gene–gene interactions [9–11]. Here, we seek to identify examples of pairwise SNP by SNP interaction effects on any of the four lipid levels in this study.

Since Bateson's first discovery in 1905 that some genes suppress the effects of other genes [12], researchers have been investigating the effect of epistasis to better understand the complex relationships between genotypes and phenotypes. Studies of model organisms suggested epistasis or gene–gene interactions are a common phenomenon [13–16], and a number of gene–gene interactions have been reported in gene mapping studies in animals, plants, and other model organisms [17–20]. However, gene–gene interactions have proven difficult to find in humans [21, 22], mainly due to low statistical power caused by the small effect size, the low minor genotype frequency of the multiple-SNP combinations, the large combinatorial number of interaction tests required [14, 23], and the lack of control over environmental conditions. Hence, to improve power and enable detection of gene–gene interactions in human GWAS, many approaches have been developed to prioritize candidate genes or SNPs using biological knowledge from established GWAS hits [6, 24], protein–protein interactions (PPIs) [25, 26], and pathway information [27].

Tests of gene–gene interactions are not as powerful as tests of single-marker association, so a judicious strategy is essential for successful epistasis analyses in human GWAS [11, 14, 15]. One fundamental limitation to the analysis of epistasis is that the statistical power of each particular test can be easily eroded by the performance of vast numbers of pairwise or higher-order interaction tests. In order to achieve similar success for interaction analyses as one obtains with single-marker tests in GWAS, we are limited to performing a similar number of tests (~1 million), assuming the nominal per-test power to detect interaction is the same as that of a single-marker association test. This limitation prevents us from conducting an inclusive all-by-all interaction analysis in current human GWAS, which typically examines millions of SNPs. If one were to attempt all-by-all epistasis tests, the result is such reduced power that only the most exceptionally strong interactions could be detected. Therefore, we recommend epistasis analysis be done on a reduced number of SNPs, using prior biological knowledge as the simplest way to restrict SNP numbers. The selection of candidate genes or SNPs can be based on any factor(s) that the biologist or medical practitioner chooses. So long as the gene choice is not

based on the tests of interactions itself, winnowing down the gene set will not inflate the type I error and can increase the power when the underlying interactions are enriched between candidate genes or SNPs. Criteria such as SNP density, local linkage disequilibrium (LD), and data quality can further supplement primary knowledge of gene function, pathway position, connectedness to networks, etc. in selecting genes.

We illustrate the biological knowledge-driven approach with examples from analyses of large consortium cohort studies of cardiovascular disease [6]. We tested for pairwise SNP by SNP epistatic interactions affecting the level of four lipids, TC, LDL-C, TG, and HDL-C, based on prior knowledge of published GWAS hits, PPIs, and pathway information. Based on GWAS hits, we detected an interaction between *HMGCR* and a locus upstream of *LIPC* in their effect on HDL-C levels in the discovery data set from the Atherosclerosis Risk in Communities (ARIC) study. Using a locus-based replication procedure, we validated this interaction in cohorts from the Framingham Heart Study (FHS) and from the Multi-Ethnic Study of Atherosclerosis (MESA). In summary, a biological, knowledge-driven approach might be crucial to the detection of epistatic interactions underlying complex traits and diseases in human GWAS.

2 Methods and Results

2.1 Descriptions of GWAS Data

Data sets from three GWAS were considered here, the Atherosclerosis Risk in Communities (ARIC) study, the Framingham Heart Study (FHS), and the Multi-Ethnic Study of Atherosclerosis (MESA). The ARIC study is a multicenter prospective investigation of atherosclerotic disease [28]. This analysis included 9,713 European Americans (EAs) in the discovery study and 3,207 African American (AA) individuals in the validation study. The FHS is a prospective cohort study to evaluate cardiovascular disease (CVD) risk factors, which has been described in detail previously [29]. This analysis included 6,575 EA subjects in the validation study, while accounting for familial relatedness (*see* Subheading 2.4 below). MESA is a prospective cohort study of individuals aged 45–84 years without clinical CVD recruited from 6 US centers [30], which is designed to study the characteristics of subclinical CVD and its progression. Participants of MESA self-reported their race or ethnicity group as EA, AA, Chinese American, or Hispanic American (HA). In total 2,685 EA, 2,588 AA, and 2,174 HA individuals were included in the validation. All of the included subjects from the three GWAS studies have both genotypic and phenotypic (lipid) measurements available, as described in the following sections.

2.2 Genotype Data and Quality Control (QC)

Genotyping of samples was obtained from the ARIC and MESA studies by Affymetrix 6.0 SNP arrays. Affymetrix 6.0 SNP array genotyping of MESA samples and Affymetrix 500K SNP array genotyping of FHS samples were obtained from dbGaP (MESA SHARe, downloaded in May 2011; Framingham Cohort, downloaded in April 2010) [31]. Standard quality control (QC) filters were applied across each of the samples and SNPs, including (1) exclusion of subjects with >10 % missing data; (2) removing SNPs with call rates <90 %; (3) removing SNPs with minor allele frequencies (MAF) ≤ 1 %; and (4) removing SNPs with Hardy–Weinberg Equilibrium (HWE) test with $P < 10^{-6}$. For the SNP pairs to be tested, we also required (1) sample size of each of the nine two-SNP combinations greater than 20 in the discovery analysis and greater than 10 in the validation study; and (2) LD measure of $r^2 < 0.1$ between the two candidate SNPs.

When necessary, untyped SNPs were imputed using IMPUTE2 [32] with HapMap3 [33] and 1,000 Genomes [34] reference haplotypes, which resulted in about the same set of SNPs across the three studies. No imputation was performed in MESA HA samples due to lack of appropriate reference haplotype panels. SNPs with information scores <0.6 were excluded, and each genotype was imputed to be the genotype with the highest posterior probability. When the highest posterior probability is less than 0.8, the genotype was treated as missing.

2.3 Phenotype Data and Covariates

Four quantitative lipid measurements, total cholesterol (TC), LDL cholesterol (LDL-C), triglyceride (TG), and HDL cholesterol (HDL-C), were considered in the analysis. Each lipid level is measured at multiple time points and the average level per individual was used in all studies. A log transformation was applied to TG levels to normalize its distribution because of the skewness in the original distribution. Individuals self-reported to be taking lipid-lowering medications were excluded. Sex, age, age squared, body mass index (BMI) were included as covariates in all analyses. The average values for age, age squared, and BMI were also used whenever multiple measurements were available. Plate number is also included as a covariate factor in the ARIC data due to its correlation with some of the lipid levels (known as “plate effect”).

2.4 Test of Statistical Interactions

Statistical interactions between pairs of SNPs were tested on a quantitative trait. For each individual, let Y denote the trait of interest and G_i denote the genotype of the i th SNP ($i = 1, 2$). G_i is the number of copies of the reference allele, thus with possible values 0, 1, or 2. Two indicator variables, x_i and z_i , were defined for each of the two SNPs as,

$$x_i = \begin{cases} 1, & G_i = 0 \\ 0, & G_i = 1 \\ -1 & G_i = 2 \end{cases} \quad z_i = \begin{cases} -0.5, & G_i = 0 \\ 0.5, & G_i = 1 \\ -0.5, & G_i = 2 \end{cases}$$

Two linear models were fitted and compared for testing of interaction. The first model (M1) included additive and dominance effect terms for each of the two SNPs without including any interaction effects between the two SNPs. The second model (M2), on top of M1, allows for four classic forms of epistatic interactions (additive \times additive, additive \times dominance, dominance \times additive, and dominance \times dominance), as follows:

$$Y = Z_0\beta_0 + x_1a_1 + z_1d_1 + x_2a_2 + z_2d_2 + \varepsilon \quad (M_1)$$

$$Y = Z_0\beta_0 + x_1a_1 + z_1d_1 + x_2a_2 + z_2d_2 + x_1x_2i_{aa} + x_1z_2i_{ad} + z_1x_2i_{da} + z_1z_2i_{dd} + \varepsilon \quad (M_2)$$

Here, β_0 is a vector of the intercept and possible nongenetic covariates; a_i and d_i are the additive and dominance effects of the i th SNP; and i_{aa} , i_{ad} , i_{da} , and i_{dd} denote the four classic interaction effects between the two SNPs. The existence of an epistatic interaction of any combination of the four types was tested by an F -test comparing M1 and M2 with four degrees of freedom [6, 35]. This classical test of statistical interactions is similar to the “--epistasis” option in PLINK [36], except that only additive effects and additive \times additive interactions are considered, such that an F -test with one degree of freedom is performed in PLINK. We note that the power of this test is strongly impacted by the marginal SNP allele frequencies.

Potential population stratification was corrected using the principal component (PC) approach and the familial relationship in FHS was accounted for using a mixed model approach [22, 37]. PC analysis was conducted using EIGENSOFT [37] and the top ten PCs were included in the analysis as covariates to account for population stratification in ARIC and MESA samples. For FHS, a mixed model approach was first applied to account for familial relatedness and then pairwise interaction was tested on the residuals from the mixed models [22].

2.5 Locus-Based Validation of Interactions

We sought to validate (replicate) the interactions detected in the discovery study using data from FHS, MESA, and another AA sample from the ARIC study. “Validate” rather than “replicate” was used here because linked and proximate SNPs were included in the validation in addition to the original SNPs, as follows. For a significant interaction between two SNPs (e.g., A and B) in the discovery study, we performed the following possible stages of tests in the validation study: (1) Test for interaction directly between SNP A and SNP B; (2) If the interaction is not significant in stage (i), test for interactions between SNP A and each SNP less than 200 kb away from SNP B, and, similarly, between SNP B and each SNP surrounding SNP A; (3) If no test in stage (ii) is significant following multiple-testing correction, test for interactions between each SNP less than 100 kb away from A and each SNP less than 100 kb away from B. Assume there are n_1 and n_2 SNPs,

respectively, surrounding SNPs A and B and the number of statistical tests to be conducted is 1, $n_1 + n_2$, and n^2 in the three validation stages, respectively. To reduce the number of tests and the cost of multiple-testing correction on power, the validation process proceeds sequentially and stops at any stage where significant results were found after multiple-testing correction.

2.6 Prioritize SNP Pairs Using Biological Knowledge

Though only pairwise interactions are considered, the total number of possible interaction tests across 2.5 million SNPs is still huge, at more than 3×10^{12} tests. Due to the severe reduction in power entailed by stringent multiple-testing correction for such a large number of tests, it is crucial to restrict the total number of tests for the whole study. Therefore, through the following three strategies, we aimed to reduce the total number of interaction tests and to enrich interaction signals in the tests considered.

2.6.1 Lipid GWAS Results

Recently, 95 genetic loci were associated with TC, LDL-C, TG, or HDL-C in a GWAS meta-analysis [7]. We considered 125 SNPs, previously associated with lipid levels [7], in the 95 loci and tested each pair of SNPs on TC, LDL-C, TG, and HDL-C in the discovery sample of 9,713 EAs from the ARIC study. Using this approach, we tested $\sim 7,748$ pairwise interactions for each trait. We identified one significant interaction affecting LDL-C levels and one interaction on HDL-C. The interaction on LDL-C was between rs2247056 and rs1030431 ($P_c = 0.03$ after Bonferroni correction). Both rs2247056 and rs1030431 were marginally associated with LDL-C, TG, and TC [7, 38]. We then performed fine mapping analyses on the loci surrounding the two SNPs to find the most significant interaction between rs2853928 and rs1993453 ($P_c = 0.01$). We tried to validate the interaction on LDL-C in additional replication samples from ARIC, FHS, and MESA, but had no success.

The interaction on HDL-C was between rs12916 and rs1532085 ($P_c = 0.008$) in the discovery study and was successfully validated in the replication samples (Table 1). To further explore the interaction between the two loci, we tested interactions between each SNP surrounding rs12916 and each SNP surrounding rs1532085 within 100 kb. While many of these SNP pairs show significant interactions due to LD, the strongest signal was between rs3846662 and rs2043085 ($P_c = 0.002$). SNP rs3846662, located in the intron of gene *HMGCR*, has been previously associated with TC and LDL-C [7, 39], but not with HDL-C. Rs3846662 has also been associated with a 2.2-fold change in *HMGCR* expression in vitro [40] and alternative splicing of exon 13 [41]. The other SNP, rs2043085, has been found to be marginally associated with HDL-C [7]. Rs2043085 is located upstream of *LIPC* and is associated with the expression of the gene [6]. On top of marginal effects, interaction between the two SNPs affects HDL-C twice as much as the effect of *LIPC* alone.

Table 1
Significant interactions affecting HDL-C levels validated in multiple population cohorts (Regenerated from ref. 6)

Test stage	SNP 1					SNP 2					P_e^b
	Cohort	rsID	Chr	Pos ^a	Gene	rsID	Chr	Pos ^a	Gene		
Discovery	ARIC EA	rs12916	5	74656539	HMGCR (3' UTR)	rs1532085	15	58683366	40.8 k U LIPC	0.008	
Fine mapping	ARIC EA	rs3846662	5	74651084	HMGCR (Intron)	rs2043085	15	58680954	43.2 k U LIPC	0.002	
Validation	MESA EA	rs3846662	5	74651084	HMGCR (Intron)	rs1973688	15	58582540	141.6 k U LIPC	0.006	
Validation	FHS EA	rs55727654	5	74651864	HMGCR (Intron)	rs473422	15	58666341	57.8 k U LIPC	0.002	
Validation	MESA HA	rs1423527	5	74602699	30.3 k U HMGCR	rs7163280	15	58718340	5.8 k U LIPC	0.04	
Validation	ARIC AA	rs3761743	5	74685520	27.6 k D HMGCR	rs567838	15	58736623	LIPC (Intron)	0.004	

^U upstream of, ^D downstream of

^aBuild 37.1 (GRCh37)

^b*P*-value after Bonferroni correction

On average, individuals with the TT genotype at rs2043085 show an increase of 2.63 mg/dl in HDL-C; subjects with TT at rs2043085 and AA at rs3846662 exhibit an average increase of 5.72 mg/dl. The linear model with these two SNPs and their interaction has an r^2 value of 0.8 %, meaning that these two SNPs and their interaction combined to explain 0.8 % of phenotypic variation in HDL-C. Using the locus-based validation procedure, the interaction between rs3846662 and rs2043085 on HDL-C was successfully validated in stage (ii) for MESA EA samples and in stage (iii) for FHS EA, MESA HA, and ARIC AA cohorts (Table 1). It did not validate after multiple-testing correction for the MESA AA sample, due to smaller sample size.

2.6.2 Protein–Protein Interactions (PPIs)

Over 3,000 high-confidence human PPIs were carefully assembled [42]. For each gene pair indicated by a PPI, we exhaustively tested all pairwise interactions between each SNP in the first gene and each SNP in the second gene. To map SNPs to genes, gene information (hg18) was obtained from the UCSC genome browser [43] and we considered all SNPs located between 5 kb upstream and downstream of a gene. Suppose there are n_1 and n_2 SNPs in the first and second genes, respectively, the number of possible pairwise interactions is $n_1 \times n_2$. For each of the four lipid traits, we performed ~6 million pairwise interaction tests according to the 3,000 PPIs, which resulted in no significant interactions following multiple-testing correction.

2.6.3 Lipid Pathway Information

We hypothesized that possible gene–gene interactions are enriched between genes in the lipid-related pathways. To test this hypothesis, we used the metabolic pathway of lipids and lipoproteins as an example. There are a total of 228 genes in this pathway [44] and 12,716 SNPs are mapped to the 228 genes. We tested all pairwise interactions among the 12,716 SNPs, resulting in a total of ~27 million interaction tests for each lipid trait. Hence, it is not surprising that we found nothing significant after multiple-testing correction. However, there is a deviation in the QQ plot of the P values for interactions underlying TC levels. The interaction between rs4804546 and rs914196 is the strongest, though not significant following correction for the ~27 million tests ($P_c=0.14$). The interaction is between two genes, *CARM1* and *AGPAT3*, from the metabolism of lipids and lipoproteins pathway. Gene *AGPAT3* has been associated with phospholipid levels [45], and *CARM1* has not been associated with any lipid levels.

3 Notes

To ensure accuracy and power, quality control process is of crucial importance in genetic association studies, as well as in studies of epistasis [46, 47]. Non-normality distribution and outliers of the quantitative phenotype can potentially lead to false positive results

with very small P values, particularly when individuals carrying the minor multi-SNP genotype incidentally are also outliers of the phenotype. When both outliers and low-frequency SNPs exist in the data, the chance of false positives can be inflated. Therefore, a minor multi-SNP genotype frequency filter (20 in discovery and 10 in validation in this study) is necessary when testing for interactions, much like the MAF filtering in a typical GWAS quality control.

While the interaction we identified was replicated in multiethnic populations, it still has a moderate effect size, which is approximately the same magnitude as the marginal effect discovered in GWAS for lipid traits. If this is the case for general epistatic interactions, we will have a lower power for detecting epistasis than marginal associations, which also explains why so few interactions have been detected and replicated in human GWAS. Moreover, unfortunately, epistatic interactions with such small effect sizes may only explain a minor proportion of the missing heritability, unless as some argue, there are a great many weakly interacting pairs.

Note that the interaction we detected in this study was validated in part by proximate SNPs, thus indicating the power of integrating information from genomic regions surrounding target SNPs, like a gene-based test for marginal associations. Recently, a series of gene-based interaction testing methods have been developed in the literature [48–52], which can be employed to increase power of detecting and replicating gene–gene interactions.

In summary, we detected significant interactions after multiple-testing correction only in <10k tests guided by GWAS, but found nothing significant in 6 and 27 million tests using PPI and pathway information, respectively. By noticing that the study is already powerful with a sample size of over 10,000, multiple measurements of phenotypes, and a genome-wide one million SNPs, we only afford to perform <10k tests to be able to identify the significant gene–gene interactions from false positives. In conclusion, a small-scale, biological, knowledge-driven study with higher enrichment of putative signals may hold the key to identifying gene–gene interactions underlying complex diseases or traits in current and future association studies.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106:9362–9367
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA et al (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
3. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 10: 241–251
4. Maher B (2008) Personal genomes: the case of the missing heritability. *Nature* 456:18–21
5. Eichler EE, Flint J, Gibson G, Kong A, Leal SM et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11:446–450

6. Ma L, Ballantyne CM, Belmont JW, Keinan A, Brautbar A (2012) Interaction between SNPs in the RXRA and near ANGPTL3 gene region inhibit apolipoprotein B reduction following statin-fenofibrate acid therapy in individuals with mixed dyslipidemia. *J Lipid Res* 53(11): 2425–2428
7. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466:707–713
8. Asselbergs FW, Guo YR, van Iperen EPA, Sivapalaratnam S, Tragante V et al (2012) Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am J Hum Genet* 91:823–838
9. Cheverud JM, Routman EJ (1995) Epistasis and its contribution to genetic variance components. *Genetics* 139:1455–1461
10. Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39:859–882
11. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci* 109:1193–1198
12. Bateson W, Saunders ER, Punnett RC, Hurst CC (eds) (1905) Reports to the Evolution Committee of the Royal Society, report II. Harrison and Sons, London
13. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* 5:618–625
14. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 10:392–404
15. Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. *Am J Hum Genet* 85:309–320
16. Gao H, Granka JM, Feldman MW (2010) On the classification of epistatic interactions. *Genetics* 184:827–837
17. Shimomura K, Low-Zeddies SS, King DP, Steeves TDL, Whiteley A et al (2001) Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res* 11:959–980
18. Carlborg Ö, Kerje S, Schütz K, Jacobsson L, Jensen P et al (2003) A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res* 13:413–421
19. Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proc Natl Acad Sci U S A* 101:15670
20. Clark AG, Doane WW (1984) Interactions between the amylase and adipose chromosomal regions of *Drosophila melanogaster*. *Evolution* 957–982
21. Ma L, Dvorkin D, Garbe J, Da Y (2007) Genome-wide analysis of single-locus and epistasis single-nucleotide polymorphism effects on anti-cyclic citrullinated peptide as a measure of rheumatoid arthritis. *BMC Proc* 1:S127
22. Ma L, Yang J, Runesha HB, Tanaka T, Ferrucci L et al (2010) Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham Heart Study data. *BMC Med Genet* 11:55
23. Ma L, Runesha HB, Dvorkin D, Garbe JR, Da Y (2008) Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. *BMC Bioinformatics* 9:315
24. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Locus* 2:0.0
25. Jia P, Zheng S, Long J, Zheng W, Zhao Z (2011) DmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics* 27:95
26. Sun YV, Kardia SLR (2010) Identification of epistatic effects using a protein-protein interaction database. *Hum Mol Genet* 19:4345
27. Wu X, Dong H, Luo L, Zhu Y, Peng G et al (2010) A novel statistic for genome-wide interaction analysis. *PLoS Genet* 6:e1001131
28. Williams OD (1989) The atherosclerosis risk in communities (ARIC) study – design and objectives. *Am J Epidemiol* 129:687–702
29. Dawber TR, Meadors GF, Moore FE (1951) Epidemiological approaches to heart disease: the Framingham study. *Am J Public Health Nations Health* 41:279–286
30. Bild DE, Bluemke DA, Burke GL, Detrano R, Roux AVD et al (2002) Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol* 156:871–881
31. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K et al (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 39:1181–1186
32. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5: e1000529

33. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF et al (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58
34. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A et al (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073
35. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11:2463–2468
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
37. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
38. Haas BE, Horvath S, Pietilainen KH, Cantor RM, Nikkola E et al (2012) Adipose co-expression networks across Finns and Mexicans identify novel triglyceride-associated genes. *BMC Med Genomics* 5:61. doi:10.1186/1755-8794-1185-1161
39. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM et al (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 41: 47–55
40. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R et al (2008) Common SNPs in HMGR in Micronesians and Whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arterioscler Thromb Vasc Biol* 28:U2078–U2332
41. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R et al (2008) Common SNPs in HMGR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arterioscler Thromb Vasc Biol* 28:2078–2084
42. Das J, Yu H (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 6:92
43. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH et al (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006
44. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D et al (2009) Reactome knowledge-base of human biological pathways and processes. *Nucleic Acids Res* 37:D619–D622
45. Lemaitre RN, Tanaka T, Tang WH, Manichaikul A, Foy M et al (2011) Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE consortium. *PLoS Genet* 7
46. Lambert CG, Black LJ (2012) Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics* 13:195–203
47. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P et al (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678
48. He J, Wang K, Edmondson AC, Rader DJ, Li C et al (2011) Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur J Hum Genet* 19: 164–172
49. Oh S, Lee J, Kwon M-S, Weir B, Ha K et al (2012) A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR. *BMC Bioinformatics* 13:S5
50. Ma L, Clark AG, Keinan A (2013) Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet* 9:e1003321
51. Li SY, Cui YH (2012) Gene-centric gene-gene interaction: a model-based Kernel machine method. *Ann Appl Stat* 6:1134–1161
52. Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C (2012) Multivariate detection of gene-gene interactions. *Genet Epidemiol* 36:622–630