HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

Methods Mol Biol. 2015 ; 1253: 35–45. doi:10.1007/978-1-4939-2155-3_3.

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

Li Ma[1,*], **Alon Keinan**[1], and **Andrew G. Clark**[1,2]

[1]Department of Biological Statistics and Computational Biology, Cornell University

[2]Department of Molecular Biology and Genetics, Cornell University

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 the 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, and pathway information. Using published GWAS results 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 EA cohorts from the Framingham Heart Study and the Multi-Ethnic Study of Atherosclerosis (MESA). We also validated the interaction in the ARIC African American sample and in the Hispanic American sample from MESA. Both *HMGCR* and *LIPC* are involved in the metabolism of lipid and lipoproteins, and *LIPC* itself has been associated with HDL-C. Moreover, the interaction affects HDL-C twice as much as the marginal effect of *LIPC*, with the effect of the two genes and their interaction combined explaining 0.8% of the variation in HDL. These results suggest the potential of the biological knowledge-driven approach to detect epistatic interactions in human GWAS, which may hold the key to exploring the role gene-gene interactions play in connecting genotype and phenotype in current and future genetic association studies.

Keywords

Epistasis; Gene-Gene Interaction; Biological Knowledge; GWAS; Lipid

Introduction

As of February 2013, over 1,489 publications and 8,271 single nucleotide polymorphisms (SNPs) have been collected in the catalog of genome-wide association studies (GWAS) [1]. Though these SNPs are significantly associated with human diseases and traits, most of them have a small effect and in aggregate account for a low proportion of the heritable variance

*Current address: Department of Animal and Avian Sciences, University of Maryland, College Park

[2,3,4,5,6]. 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, several meta-analyses of many GWAS, with a combined sample size of up to 100,000, have detected hundreds of loci associated with the levels of the 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 such missing heritability, including rare and structural variants, gene-environment interactions, epigenetics, and complex inheritance [2,3,4,5]. The missing heritability may also be partially attributed to epistatic or gene-gene interactions [9,10,11], and 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 in order to better understand the complex relationship between genotype and phenotype. Studies of model organisms suggested epistasis or gene-gene interactions to be a common phenomenon [13,14,15,16], and a number of gene-gene interactions have been reported in gene mapping studies in animals, plants, and other model organisms [17,18,19,20]. 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 of environmental conditions. Hence, in order to improve the power of 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 a successful epistasis analysis 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 a similar success for interaction analysis 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 the 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 analysis 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.

Methods and Results

Descriptions of GWAS data

Three GWAS data sets were considered in this study, 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 multi-center prospective investigation of atherosclerotic disease [28]. This analysis included 9,713 European Americans (EA) 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 their familial relatedness (see **Test of statistical interactions** 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 genotypes and phenotypic (lipid) measurements available, as described in the following sections.

Genotype data

Genotyping of samples was obtained from the ARIC study and MESA by Affymetrix 6.0 SNP array. 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 was imputed using IMPUTE2 [32] with HapMap3 [33] and 1000 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 the lack of appropriate reference haplotype panels. SNPs with information score < 0.6 were excluded and each genotype was imputed with the genotype with the highest posterior probability. When the highest posterior probability is less than 0.8, the genotype was treated as missing.

Phenotype data

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”).

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 ith 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,

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

Two linear models were fitted and compared for testing of interaction. The first model (M_1) included additive and dominance effect terms for each of the two SNPs without including any interaction effects between the two SNPs; The second model (M_2), on top of M_1, allows for four classic forms of epistatic interactions (additive × additive, additive × dominance, dominance × additive, and dominance × 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 non-genetic covariates. a_i and d_i are the additive and dominance effects of the ith SNP. 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 M_1 and M_2 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 × additive interaction are considered, such that an *F*-test with one degree of freedom is performed in PLINK.

Potential population stratification was corrected by principal component (PC) approach and familial relationship in FHS was accounted for using a mixed model approach [22,37]. PC analysis was conducted using EIGENSOFT [37] and the top 10 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].

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 perform the following possible stages of tests in the validation study: (i) Test for interaction directly between SNP A and SNP B; (ii) If the interaction is not significant in Stage (i), test for interaction 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; (iii) If no test in Stage (ii) is significant following multiple-testing correction, test for interaction between each SNP less 100kb away from A and each SNP less than 100kb away from B. Assume there are n_1 and n_2 SNPs respectively surrounding SNPs A and B, 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.

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 fewer number of tests considered.

1. Lipid GWAS results—In total, 95 genetic loci have been recently reported to be associated with TC, LDL-C, TG, or HDL-C in a GWAS meta-analysis [7]. We considered 125 SNPs 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. The 125 SNPs have been previously associated with any of the four lipid levels [7]. Using this approach, we tested ~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 analysis in the two loci surrounding the two SNPs to find the most significant interaction to be between rs2853928 and rs1993453 (P_c = 0.01). We tried to validate this 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 interaction between each SNP surrounding rs12916 and each SNP surrounding rs1532085 within a 100 kb distance. 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 shown in vitro to be associated with a 2.2-fold change in *HMGCR* expression [40] and affect 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 associated with the expression of the gene [6]. On top of the marginal effects, the interaction between the two SNPs affects HDL-C twice as much as the effect of *LIPC* alone: While on average individuals with TT genotype at rs2043085 show an increase of 2.63 mg/dl in HDL-C, subject with TT at rs2043085 and AA at rs3846662 exhibits 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 explain a 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 significantly after multiple-testing correction for the MESA AA sample with a smaller sample size.

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 one. Suppose n_1 and n_2 are the numbers of SNPs in the first and second genes respectively, the number of possible pairwise interactions is $n_1 \times n_2$. 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 5 kb downstream of a gene. For each of the four lipid traits, we performed ~6 million pairwise interaction tests according to the ~3,000 PPIs, which, however, resulted in no significant interactions following multiple-testing correction.

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 metabolism of lipids and lipoproteins pathway 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 the level of phospholipid [45], and *CARM1* has not been found to be associated with any lipid levels.

Notes

To ensure accuracy and power, quality control is of crucial importance in genetic association studies, as well as in epistasis analysis [46,47]. Non-normality and outliers in the distribution of the phenotypes can potentially lead to false positive interactions with very small *P* values, particularly when the individuals carrying the minor multi-SNP genotype incidentally are also outliers in phenotype. When both outliers and low-frequency SNPs exist in the data, the chance of false positives will be greatly increased. 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 typical GWAS.

While the interaction we identified was replicated in multi-ethnic populations, it has a small effect size, which is about the same magnitude as the marginal effect discovered in GWAS for complex diseases and traits. If this is the case for general epistatic interactions, we will have even lower power for detecting epistasis than marginal associations, which also explains why so few interactions have been detected and replicated in human GWAS.

Note that the interaction we detected in this study was validated in part by proximate SNPs, thus indicating the power of integrating information from a genomic region surrounding the 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 [48,49,50,51,52], which can be employed to increase power of detecting and replicating interactions.

In this study, we detected significant interactions after multiple-testing correction only in < 10k tests guided by GWAS results, but found nothing significant in 6 million and 27 million tests respectively using PPI and pathway information. By noticing that the discovery study is already powerful with a sample size of ~10k, multiple measurements of the phenotypic values, and a genome-wide 1 million SNP genotyping, we only afford to perform <10k tests. 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 human association studies.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]
3. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009; 10:241–251. [PubMed: 19293820]
4. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456:18–21. [PubMed: 18987709]

5. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]
6. Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, et al. Knowledge-Driven Analysis Identifies a Gene-Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. PLoS genetics. 2012:8.
7. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
8. Asselbergs FW, Guo YR, van Iperen EPA, Sivapalaratnam S, Tragante V, et al. Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci. American Journal of Human Genetics. 2012; 91:823–838. [PubMed: 23063622]
9. Cheverud JM, Routman EJ. Epistasis and its contribution to genetic variance components. Genetics. 1995; 139:1455–1461. [PubMed: 7768453]
10. Cockerham CC. An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present. Genetics. 1954; 39:859–882. [PubMed: 17247525]
11. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences. 2012; 109:1193–1198.
12. Bateson, W.; SE, R.; PR, C.; HC, C., editors. Reports to the Evolution Committee of the Royal Society, Report II. Harrison and Sons; London, UK: 1905.
13. Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? Nature Reviews Genetics. 2004; 5:618–U614.
14. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics. 2009; 10:392–404.
15. Moore JH, Williams SM. Epistasis and its implications for personal genetics. American Journal of Human Genetics. 2009; 85:309–320. [PubMed: 19733727]
16. Gao H, Granka JM, Feldman MW. On the Classification of Epistatic Interactions. Genetics. 2010; 184:827–U351. [PubMed: 20026678]
17. Shimomura K, Low-Zeddies SS, King DP, Steeves TDL, Whiteley A, et al. Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. Genome research. 2001; 11
18. Carlborg Ö , Kerje S, Schütz K, Jacobsson L, Jensen P, et al. A global search reveals epistatic interaction between QTL for early growth in the chicken. Genome research. 2003; 13:413–421. [PubMed: 12618372]
19. Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:15670. [PubMed: 15505218]
20. Clark AG, Doane WW. Interactions between the Amylase and Adipose chromosomal regions of Drosophila melanogaster. Evolution. 1984:957–982.
21. Ma L, Dvorkin D, Garbe J, Da Y. Genome-wide analysis of single-locus and epistasis single-nucleotide polymorphism effects on anti-cyclic citrullinated peptide as a measure of rheumatoid arthritis. BMC Proceedings. 2007; 1:S127. [PubMed: 18466469]
22. Ma L, Yang J, Runesha HB, Tanaka T, Ferrucci L, et al. Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham Heart Study data. BMC Medical Genetics. 2010; 11:55. [PubMed: 20370913]
23. Ma L, Runesha HB, Dvorkin D, Garbe JR, Da Y. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. BMC bioinformatics. 2008; 9:315. [PubMed: 18644146]
24. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Locus. 2005; 2:0.0.
25. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics. 2011; 27:95. [PubMed: 21045073]

26. Sun YV, Kardia SLR. Identification of epistatic effects using a protein-protein interaction database. Human Molecular Genetics. 2010; 19:4345. [PubMed: 20736252]
27. Wu X, Dong H, Luo L, Zhu Y, Peng G, et al. A Novel Statistic for Genome-Wide Interaction Analysis. PLoS genetics. 2010; 6:e1001131. [PubMed: 20885795]
28. Williams OD. The Atherosclerosis Risk in Communities (ARIC) Study - Design and Objectives. American Journal of Epidemiology. 1989; 129:687–702. [PubMed: 2646917]
29. Dawber TR, Meadors GF, Moore FE. Epidemiological Approaches to Heart Disease: The Framingham Study. American Journal of Public Health and the Nations Health. 1951; 41:279–286.
30. Bild DE, Bluemke DA, Burke GL, Detrano R, Roux AVD, et al. Multi-ethnic study of atherosclerosis: Objectives and design. American Journal of Epidemiology. 2002; 156:871–881. [PubMed: 12397006]
31. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. The NCBI dbGaP database of genotypes and phenotypes. Nature genetics. 2007; 39:1181–1186. [PubMed: 17898773]
32. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009; 5:e1000529. [PubMed: 19543373]
33. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
34. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
35. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics. 2002; 11:2463–2468. [PubMed: 12351582]
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81:559–575. [PubMed: 17701901]
37. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38:904–909. [PubMed: 16862161]
38. Haas BE, Horvath S, Pietilainen KH, Cantor RM, Nikkola E, et al. Adipose Co-expression networks across Finns and Mexicans identify novel triglyceride-associated genes. BMC Medical Genomics. 2012; 5 Doi 10.1186/1755-8794-1185-1161.
39. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet. 2009; 41:47–55. [PubMed: 19060911]
40. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R, et al. Common SNPs in HMGCR in Micronesians and Whites Associated With LDL-Cholesterol Levels Affect Alternative Splicing of Exon13. Arteriosclerosis Thrombosis and Vascular Biology. 2008; 28:2078–U2332.
41. Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R, et al. Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. Arteriosclerosis, Thrombosis, and Vascular Biology. 2008; 28:2078–2084.
42. Das J, Yu HY. HINT: High-quality protein interactomes and their applications in understanding human disease. Bmc Systems Biology. 2012:6. [PubMed: 22260221]
43. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. The human genome browser at UCSC. Genome research. 2002; 12:996–1006. [PubMed: 12045153]
44. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research. 2009; 37:D619–D622. [PubMed: 18981052]
45. Lemaitre RN, Tanaka T, Tang WH, Manichaikul A, Foy M, et al. Genetic Loci Associated with Plasma Phospholipid n-3 Fatty Acids: A Meta-Analysis of Genome-Wide Association Studies from the CHARGE Consortium. PLoS genetics. 2011:7.
46. Lambert CG, Black LJ. Learning from our GWAS mistakes: from experimental design to scientific method. Biostatistics. 2012; 13:195–203. [PubMed: 22285994]

47. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]

48. He J, Wang K, Edmondson AC, Rader DJ, Li C, et al. Gene-based interaction analysis by incorporating external linkage disequilibrium information. European Journal of Human Genetics. 2011; 19:164–172. [PubMed: 20924406]

49. Oh S, Lee J, Kwon M-S, Weir B, Ha K, et al. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. BMC bioinformatics. 2012; 13:S5. [PubMed: 22901090]

50. Ma L, Clark AG, Keinan A. Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. PLoS genetics. 2013:9.

51. Li SY, Cui YH. Gene-Centric Gene-Gene Interaction: A Model-Based Kernel Machine Method. Annals of Applied Statistics. 2012; 6:1134–1161.

52. Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C. Multivariate Detection of Gene-Gene Interactions. Genetic epidemiology. 2012; 36:622–630. [PubMed: 22782518]

Table 1

Significant interactions affecting HDL-C levels validated in multiple population cohorts

(Regenerated from Ref [6])

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

[a] Build 37.1 (GRCh37)

[b] U = upstream of; D = downstream of

[c] *P*-value after Bonferroni correction